A data-driven approach for neonatal mortality rate forecasting

Elen Rodríguez[a], Elias Rodríguez[a], Luiz Nascimento[a,b], Aneirson da Silva[a] and Fernando Marins[a]

[a] *São Paulo State University (UNESP), Av. Dr. Ariberto Pereira da Cunha 333, Guaratinguetá/SP, 12516-410, Brazil*

[b] *University of Taubate (UNITAU), Estrada Municipal Dr. José Luiz Cembranelli 5.000, Taubaté/SP, 12.081-010, Brazil*

Abstract

Neonatal mortality is an important public health problem that reflects the development of a country, as well as the quality of care provided to the newborn. This article presents the development and comparison of classical models and machine learning models for time series forecasting, applied to the forecast of monthly neonatal mortality rates in the metropolitan region of Paraiba River Valley and North Coast – São Paulo State - Brazil. The database used comprised the monthly rates from January 2000 to December 2020. The models compared were Seasonal Autoregressive Integrated Moving Average, random forest, support vector machine (SVM), light gradient boosting machine, categorical boosting (CatBoost), gradient boosting (GB), extreme gradient boosting, and multilayer perceptron. The best parameters and hyperparameters of the models tested were adjusted through an exhaustive computational search. The results showed that the CatBoost, SVM, and GB models presented the lowest values in the error metrics evaluated, and the SVM model presented better precision. The forecasts of the SVM model showed a behavior very close to the actual rates, which was confirmed by the application of the paired t-test. These results corroborate that time series forecasting models can significantly contribute as a decision support tool for public health problems.

Keywords 1

Neonatal mortality, time series analysis, forecasting, data-driven models, machine learning

1. Introduction

Neonatal mortality is associated with the occurrence of newborn death in the first 28 days of life and is considered an important public health problem [1]. In this context, both the infant mortality rate and the neonatal mortality rate can be considered relevant indicators for evaluating the quality of care provided to newborns, public health, and the well-being of the population, as well as the country's development.

Globally, from 1990 to 2015, infant mortality has been decreasing, which has been mainly reflected in post-neonatal mortality, but unfortunately, the number of neonatal deaths has not been showing the same reduction [2], (Figure 1). Corroborating this finding, according to the World Health Organization (WHO), from 1990 to 2019, neonatal deaths decreased from 5.0 to 2.4 million, but only in 2019, 47% of deaths in children under 5 years old occurred in the first 28 days of life [3]. Therefore, reducing the number of neonatal deaths still remains a major challenge [4].

The risk of death of the newborn is greatest during the first hours and days of life [4], with approximately 1 million newborns worldwide dying in the first 24 hours of life and about 75% of deaths

1 IDDM-2022: 5th International Conference on Informatics & Data-Driven Medicine, November 18–20, 2022, Lyon, France

EMAIL: elen.aguirre@unesp.br (E. Rodríguez); elias.aguirre@unesp.br (E. Rodríguez); fernando.nascimento@unesp.br (L. Nascimento); aneirson.silva@unesp.br (A. da Silva); fernando.marins@unesp.br (F. Marins)

ORCID: 0000-0002-3829-4118 (E. Rodríguez); 0000-0003-1120-1708 (E. Rodríguez); 0000-0001-9793-750X (L. Nascimento); 0000-0002-2215-0734 (A. da Silva); 0000-0001-6510-9187 (F. Marins)

occurring during the first week of life [3]. In addition, it is important to emphasize that newborns are highly vulnerable, and those newborns with low birth weight, premature delivery, or with some health problems are even more vulnerable and have a higher risk of death in the neonatal period [1].

Reducing infant and neonatal mortality has become a global concern and has been reflected in the approach to the Millennium Development Goals (MDGs) and the Sustainable Development Goals (SDGs). Specifically, the MDGs sought to reduce infant mortality by 2015 [5], and Goal 3 of Health and Well-being of the SDGs aims to reduce neonatal mortality to at least 12 per 1,000 live births by 2030 [6].

In the case of Brazil, from 1996 to 2020, infant mortality rates decreased, reaching in 2015 the established target of the MDGs in reducing infant mortality [7]. However, it appears that the reduction in the number of infant deaths was mainly due to the decrease in post-neonatal mortality (between 28 and 365 days of life) [4,8], as shown in Figure 1.

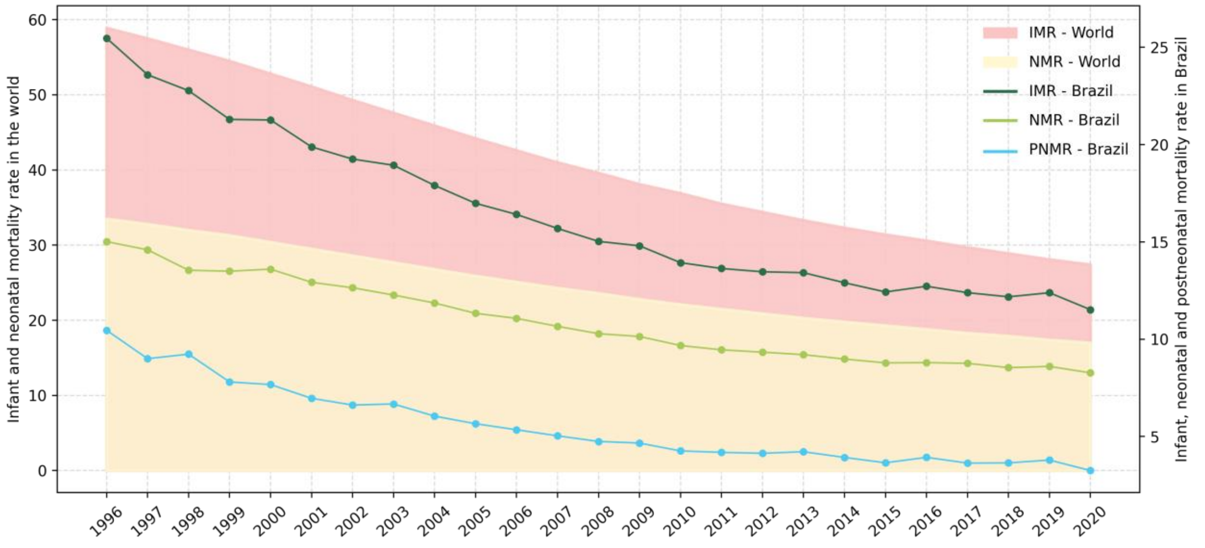

Figure 1: Infant and neonatal mortality rate in the world and infant, neonatal, and post-neonatal mortality rate in Brazil.

To improve the conditions for decision-making in public health problems, it is possible to use technological advances and, with the historical data stored, it is possible to extract useful knowledge to predict future events [9-11]. Thus, current and past knowledge can be used to model and make future predictions that can help managers to face the serious problems already mentioned [11], making it possible to identify areas of uncertainty, such as quantitative mortality predictions [12].

In this context, time series forecasting plays an important role and can be a useful tool for planning public health policies and improving the provision of health services and care [12-14].

In this way, several algorithms can be applied to the development of time series models, which can capture complex patterns in the available data [9, 11, 14]. Some of the classic time series methods are the Autoregressive Integrated Moving Average (ARIMA), and the Seasonal Autoregressive Integrated Moving Average (SARIMA) also known as Seasonal ARIMA, among others [9, 11].

On the other hand, in recent years there has been a growing interest in the application of Machine Learning (ML) in several areas of study [15], and the medical field has been no exception, and ML has been used as a decision support tool in several problems [16, 17]. In addition, ML techniques have been used for time series forecasting, which has resulted in the development of efficient models that often outperform classical models [9, 11].

It is important to highlight that in the literature there are still few works on health forecasts with time series [13]. Therefore, the general objective of this article was to present the development of forecasting models in time series, using classical statistical techniques, and ML techniques, using real data on neonatal mortality in the metropolitan region of Paraiba River Valley and North Coast – São Paulo State - Brazil.

The article is organized as follows: Section 2 describes the materials and methods adopted; Section 3 presents the results and their discussion, describing the main findings. Finally, Section 4 presents the conclusions, followed by the bibliographic references.

2. Materials and Methods

Figure 2 illustrates the methodological procedure in this time series forecast analysis, from 2000 to 2020, of the neonatal mortality rate in the metropolitan region of Paraiba River Valley and North Coast – São Paulo State - Brazil.

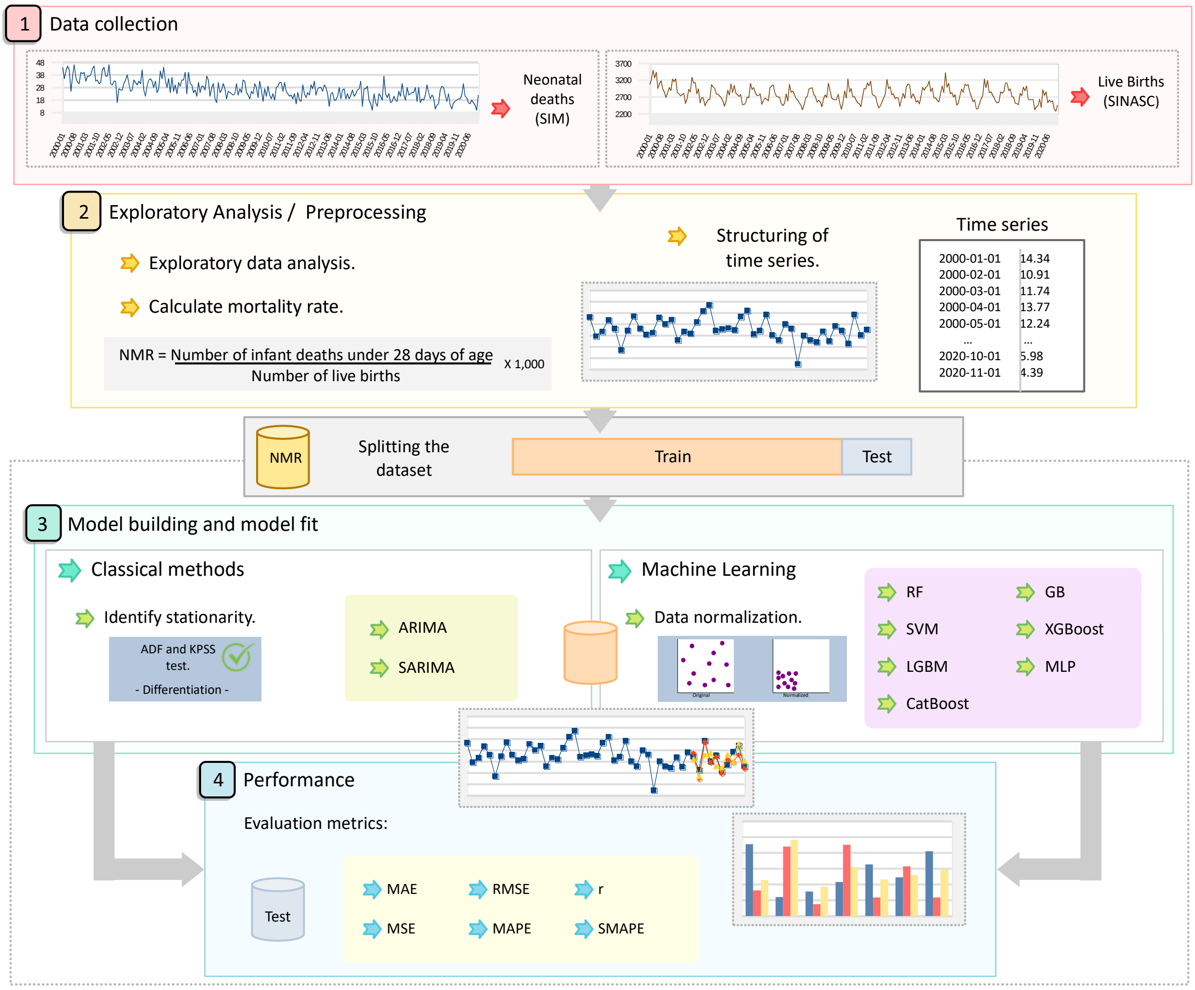

Figure 2: Steps of the methodological process.

The methodological process of this work consisted of a sequence of 4 main steps, which are described below:

- **Step 1 - Data Collection:**

  The database for this study was obtained from the Department of Informatics of the Unified Health System (DATASUS) [18], considering the records from January 2000 to December 2020 in the metropolitan region of Paraiba River Valley and North Coast – São Paulo State - Brazil. Specifically, neonatal death records were extracted from the Mortality Information System (SIM), and live birth records were obtained from the System of Information of Live Births (SINASC) [18].

- **Step 2 - Exploratory Analysis and Preprocessing:**

The collected data were analyzed using Descriptive Statistics techniques and missing data were identified. Monthly neonatal mortality rates were calculated by dividing the number of neonatal deaths by the number of live births per month, expressed per 1,000, as formulated in (1):

$$NMR = \frac{Number\ of\ infant\ deaths\ under\ 28\ days\ of\ age}{Number\ of\ live\ births} \times 1{,}000 \tag{1}$$

Then, the monthly time series was structured and graphs were built to better visualize the behavior of these data. In addition, the dataset was split in a 90:10 ratio for training and testing, considering rates from January 2000 to November 2018 for training and from December 2018 to December 2020 for testing.

- **Step 3 - Forecasting Models:**

A time series is a sequence of chronologically ordered random observations, denoted by $y = \{y_t, y_{t+1}, y_{t+2}, \ldots, y_T\}$, where y_t is an observation at instant t satisfying $1 \leq t \leq T$, where T the duration of the entire period considered. As can be seen in the available literature, classical statistical methods, as well as ML techniques, have been successfully used in time series forecasting [9, 11].

In this work, the SARIMA, Random Forest (RF), Support Vector Machine (SVM), Light Gradient Boosting Machine (LGBM), Categorical Boosting (CatBoost), Gradient Boosting (GB), eXtreme Gradient Boosting (XGBoost), and Multilayer Perceptron (MLP) models were tested, and their performances were compared in forecasting monthly neonatal mortality rates. The models were implemented with the Python v.3.9.7 programming language and in the Jupyter-notebook v.6.4.6 development environment.

The SARIMA model is an extension of the ARIMA model, being that the ARIMA(p,d,q) model includes the statistical procedures of autoregression (AR) of order p, the integration (I) that indicates the number of differences d necessary to guarantee the stationarity of the series, and the moving average (MA) of order q. In the case of the SARIMA$(p,d,q)(P,D,Q)_s$ model, the letter 's' subscript represents the seasonality period of the series, and P, D, and Q are respectively the autoregressive, differential, and moving average procedures of the seasonal part of the ARIMA process [9].

A time series is stationary when it satisfies the properties of the mean ($E(y_t)$=μ), variance ($var(y_t)$=$E(y_t - \mu)^2$=σ^2) and autocovariance (γ_k=$Cov(y_t, y_{t+k})$=$E[(y_t - \mu)(y_{t+k} - \mu)]$) do not change over time. Many of the analyzed data are not stationary series and the common practice is to transform non-stationary data into stationary data through successive differences [9], being that generally, the first difference (Δy_t=$y_t - y_{t-1}$) the first difference is enough to convert the series from non-stationary to stationary.

The Augmented Dickey–Fuller (ADF) test and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test were used to verify stationarity in the monthly neonatal mortality rate data, adopting a significance level of 5%. The null hypothesis ($\rho = 1$) of the ADF test is the existence of a unit root, that is, the series is not stationary, and the null hypothesis (H_0: $|\rho| < 1$) of the KPSS test is that there is no unit root, that is, the series is stationary [19,20].

Thus, the autoregressive (AR) process of order p is given in (2):

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \tag{2}$$

where ϕ_p is the model parameters and ε_t is the random error also known as white noise.

Equation (3) corresponds to the moving averages (MA) model of order q:

$$y_t = \theta_0 - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{3}$$

where θ_q is the parameters of the MA model.

The ARIMA model is based on the stationarity assumption, applying the combination of the AR model with the MA model in the differentiated series, as presented in (4):

$$y_t = \phi_0 + \phi_1 y_{t-1}^d + \phi_2 y_{t-2}^d + \cdots + \phi_p y_{t-p}^d + \theta_0 - \theta_1 \varepsilon_{t-1}^d - \theta_2 \varepsilon_{t-2}^d - \cdots - \theta_q \varepsilon_{t-q}^d \tag{4}$$

where y_t^d is a differentiated time series of order d [9,11,20].

In the case of the SARIMA model, equation (5) is used:

$$y_t = \phi_0 + \phi_1 y_{t-1}^d + \phi_2 y_{t-2}^d + \cdots + \phi_p y_{t-p}^d + \theta_0 - \theta_1 \varepsilon_{t-1}^d - \theta_2 \varepsilon_{t-2}^d - \cdots - \theta_q \varepsilon_{t-q}^d + \\ \Phi_0 + \Phi_1 y_{t-1}^D + \Phi_2 y_{t-2}^D + \cdots + \Phi_{Ps} y_{t-Ps}^D + \Theta_0 - \Theta_1 \varepsilon_{t-1}^D - \Theta_2 \varepsilon_{t-2}^D - \cdots - \Theta_{Qs} \varepsilon_{t-Qs}^D \tag{5}$$

where ϕ_p and θ_q are the non-seasonal parameters of the AR and MA model, Φ_{Ps} and Θ_{Qs} are the seasonal parameters of the AR and MA model, respectively, s is the seasonality period, d is the non-seasonal differentiation, and D is the differentiation of the seasonal part of the model [11, 21].

On the other hand, in the case of ML models, the original time series data from the training set were normalized using the MIN-MAX method. The MIN-MAX method transforms the data in the range [0 – 1] [10, 17, 22], as shown in (6):

$$y'_t = \frac{y_t - y_{min}}{y_{max} - y_{min}} \tag{6}$$

where y_{min} and y_{max} are, respectively, the minimum and maximum rates of the training dataset, and y'_t is the normalized value of y_t [17, 22].

For the development of supervised ML models, all normalized observations of the training series (S) were considered. Thus, the training dataset was formatted in a subsequence of observations $S = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_i, Y_i)\}$, as presented in (7):

$$X_i = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,d} \\ y_{2,2} & y_{2,3} & \cdots & y_{2,d+1} \\ y_{3,3} & y_{3,4} & \cdots & y_{3,d+2} \\ \vdots & \vdots & \cdots & \vdots \\ y_{i,d+(i-d)} & y_{i,d+(i-2)} & \cdots & y_{i,d+(i-1)} \end{bmatrix} \wedge \; y_i = \begin{bmatrix} y_{1,d+1} \\ y_{2,d+2} \\ y_{3,d+3} \\ \vdots \\ y_{i,d+i} \end{bmatrix} \tag{7}$$

where X_i is the matrix of input attributes of the temporal pattern of dimension d, and y_i is the vector of labels.

Furthermore, d is the size of the sliding window, that is, the number of previous mortality rates considered to predict y_{d+1}, which indicates the posterior value of $y_{1,d}$. In other words, ML models were trained with a finite sequence of $X \times Y$ pairs, where d mortality rates from previous months were used to forecast a future observation.

Among the ML techniques that can be used to forecast time series are neural networks such as MLP, SVM and ensemble methods such as RF, GB, LGBM, CatBoost, and XGBoost.

MLP is a neural network that has a layered structure (input layer, one or more hidden layers, and output layer), composed of neurons also known as perceptrons. From the MLP network architecture, it is possible to approximate continuous functions providing a non-linear mapping between inputs and outputs [9, 23]. The dimension of the input attributes is equal to the number of neurons in the input layer of the neural network, as well as the number of neurons in the output layer corresponds to the target variable [9]. Each neuron in the network has an activation function that will determine the perceptron output, in addition to each neuron receiving information from the neuron of the previous layer until the MLP prediction is generated [23].

SVM uses the input data to map it onto a high-dimensional non-linear hyperplane, also known as a feature space [24]. The construction of the SVM model is subject to linear constraints and the estimation of unknown parameters, such as the weight vector and the hyperplane bias [9]. The hyperplane estimation is performed using the training data and the Lagrange multiplier method to define the edges of the ideal separating hyperplane [24].

The ensemble methods were proposed to reduce the variation and improve the accuracy of the developed models [25]. A wide variety of ensemble methods have been proposed, such as the RF, GB, LGBM, CatBoost, and XGBoost methods. The construction of the models of the ensemble method are based on decision trees, but the training process for each model is different. The decision trees of the models and their variations based on gradient boosting are trained iteratively, unlike the RF models that train the decision trees in parallel [25].

- **Step 4 - Performance Metrics:**

  According to the literature, there is no single standard evaluation metric [22, 26], so, here, the model accuracies were evaluated and compared using various metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) [22], Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Error (SMAPE) [27], and Correlation Coefficient (r). The calculations of these metrics can be performed by (8)-(13):

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \tag{8}$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{10}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t} \times 100\right| \tag{11}$$

$$SMAPE = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{|y_t - \hat{y}_t|}{(|y_t| + |\hat{y}_t|)/2}\right) \tag{12}$$

$$r = \frac{n(\sum y_t\,\hat{y}_t) - (\sum y_t)(\sum \hat{y}_t)}{\sqrt{[n\sum y_t^2 - (\sum y_t)^2][n\sum \hat{y}_t{}^2 - (\sum \hat{y}_t)^2]}} \tag{13}$$

where y_t is the actual value of the mortality rate for the period $t = 1,2,3,\ldots,n$, $\bar{y}$ is the average of y, and $\hat{y}_t$ is the predicted value of observation t.

The model with the best performance was selected by evaluating the lowest values in the *MAE*, *MSE*, *RMSE*, *MAPE*, and *SMAPE* metrics, in addition to the highest value in the correlation coefficient [22, 27, 26]. In addition, the predicted rates and the actual rates were compared using the paired t-test to verify significant differences between the values of each model ($t - value < t - critical$), adopting a significance level of 5% ($p - value > 0.05$) [28].

3. Results and Discussion

In this section, the main results of the descriptive analysis of the data are presented, besides the predictions of the neonatal mortality rate of the models tested.

In the analyzed period from 2000 to 2020, the metropolitan region of Paraiba River Valley and North Coast – São Paulo State – Brazil had 697,903 live births and 6,631 neonatal deaths, with an average of 2,769 live births and 26 deaths per month. Using descriptive statistics techniques, applied to monthly neonatal mortality rates, the minimum and maximum monthly rates were obtained, which were 3.71 and 16.47, respectively. The mean monthly rate was 9.48, the standard deviation was equal to 2.54, the median was equal to 9.20, and the first and third quartiles were equal to 7.52 and 11.25 respectively.

Figure 3 presents the annual (A) and monthly (B) boxplot of the collected data (2000-2020), in which the distribution and behavior of values over time can be observed. In the annual analysis, it was observed that neonatal mortality rates had a decreasing behavior, in addition to showing a slight increase in 2017 and a variable behavior in the next three years. Likewise, in the monthly analysis, it was observed that June presented a decrease and less variation in mortality rates.

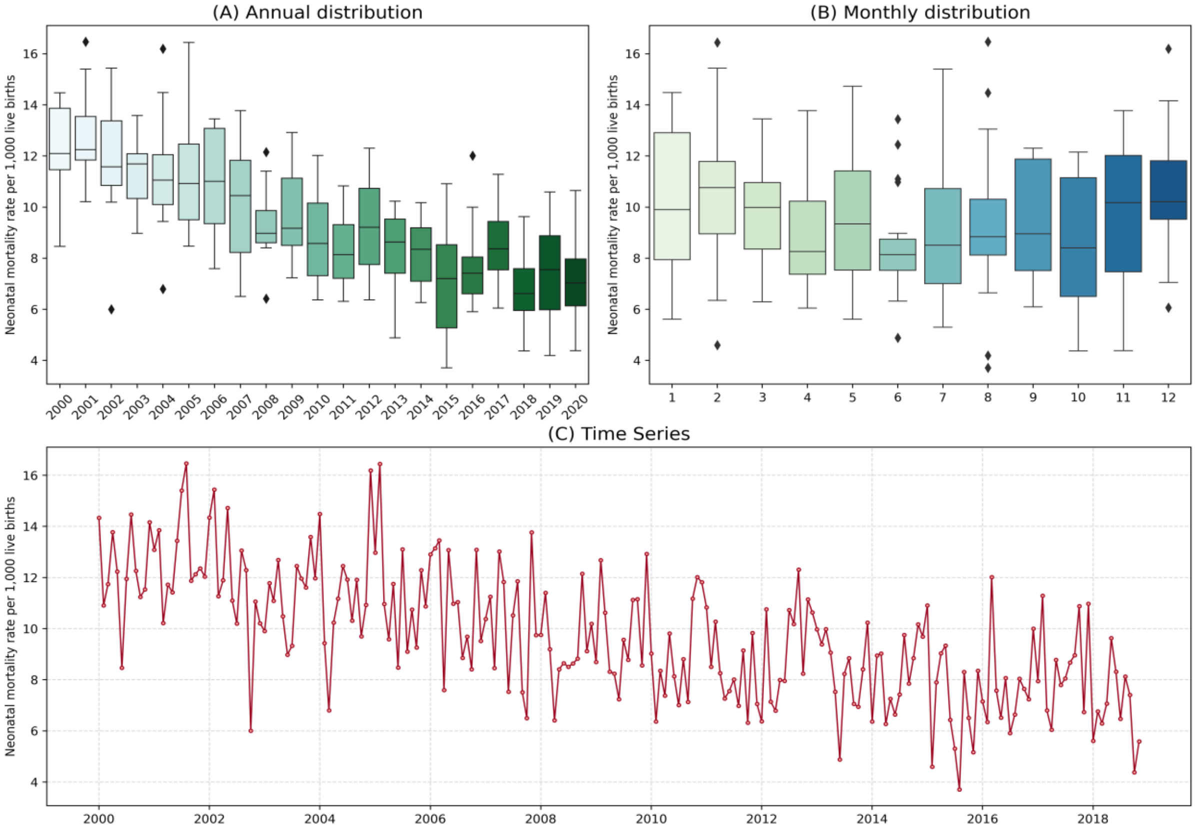

Figure 3: (A) Annual and (B) monthly boxplot of the collected data, and (C) the original time series used for the models.

Figure 3 (C) shows the time series line graph of the neonatal mortality rate for the 227 months considered for training the models, and it was possible to observe the presence of a decreasing trend and a possible variation of the mean, which may indicate the non-stationarity of the series.

Furthermore, it is known that in the case of the classical SARIMA method, the stationarity of the time series is a prerequisite [9, 11]. The collected data were analyzed and evaluated for the possible presence of non-stationarity through the ADF and KPSS tests, and these results are presented in Table 1.

Table 1

Verification of data stationarity using the Augmented Dickey–Fuller (ADF) test and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test.

Neonatal mortality rate		ADF[a]	p-*value*	KPSS[b]	p-*value*
Original time series	Test Statistic	-2.021	0.277	2.271	0.010
	CV[c] (1%)	-3.461		0.739	
	CV (5%)	-2.875		0.463	
	CV (10%)	-2.574		0.347	
1st difference	Test Statistic	-7.692	<0.001	0.041	0.100
	CV (1%)	-3.461		0.739	
	CV (5%)	-2.875		0.463	
	CV (10%)	-2.573		0.347	

(a) ADF: Augmented Dickey–Fuller; (b) KPSS: Kwiatkowski–Phillips–Schmidt–Shin; (c) CV: critical value.

The results of the ADF and KPSS tests for the analysis of the existence of a unit root in the original time series confirmed that the data are non-stationary (Table 1), and therefore, a first-order difference was applied to the series to make it stationary. The stationarity of the differentiated series was confirmed by the ADF and KPSS tests at a significance level of 5%, revealing the rejection of the null hypothesis ($ADFvalue < criticalvalue$ and $p - value < 0.05$) in the ADF test and the acceptance of the null hypothesis ($KPSSvalue < criticalvalue$ and $p - value > 0.05$) in the KPSS test.

The estimations of the best parameters of the SARIMA$(p,d,q)(P,D,Q)_s$ model were performed by applying an exhaustive computational search. Furthermore, the models with the best combinations of parameters were selected using the AIC to select those with the best fit. Table 2 presents the best parameter configuration of the SARIMA model.

Table 2
SARIMA model adjusted for neonatal mortality rates per 1,000 live births.

Model	Parameters	Std. error	p-*value*	95% CI[a]	AIC
SARIMA$(3,1,1)(1,1,1)_{12}$	ϕ_1=0.097	0.081	0.233*	[-0.062; 0.255]	944.543
	ϕ_2=0.021	0.068	0.759*	[-0.113; 0.155]	
	ϕ_3=0.084	0.080	0.298*	[-0.074; 0.241]	
	θ_1=-0.945	0.028	<0.001	[-1.005; -0.894]	
	$\Phi_{1,12}$=0.956	0.067	<0.001	[0.825; 1.087]	
	$\Theta_{1,12}$=-0.886	0.110	<0.001	[-1.102; -0.669]	

(a) CI: Confidence interval; * Not significant (p-value>0.05).

Likewise, Table 3 presents the best settings for the hyperparameters of the tested ML models. The search for the best hyperparameters was performed using the grid search method, in which the parameters selected were those that maximized the performance of the model.

Table 3
Best hyperparameter settings of machine learning models

Model	Settings
RF	Window size: 25, n_estimators: 128, min_samples_leaf: 1, min_samples_split: 2, bootstrap: False, criterion: 'squared_error', max_depth: 5
SVM	Window size: 12, C: 2.3, cache_size: 200, degree: 3, epsilon: 0.095, gamma: 'scale', kernel: 'rbf', tol: 0.01
LGBM	Window size: 26, n_estimators: 50, num_leaves: 41, boosting_type: 'dart', colsample_bytree: 1, importance_type: 'gain', learning_rate: 0.1, max_depth: -1, min_child_samples: 5, min_child_weight: 0.0001
CatBoost	Window size: 24, iterations: 100, learning_rate: 0.5, depth: 8, loss_function: 'RMSE'
GB	Window size: 25, n_estimators: 197, alpha: 0.9, criterion: 'friedman_mse', learning_rate: 0.09, loss: 'squared_error', max_depth: 3, min_samples_leaf: 1, min_samples_split: 2, min_weight_fraction_leaf: 0.0, tol: 0.0001
XGBoost	Window size: 24, n_estimators: 35, base_score: 0.65, booster: 'gbtree', learning_rate: 0.300, max_delta_step: 0, reg_lambda: 2, n_jobs: 5, num_parallel_tree: 1, max_depth: 6, min_child_weight: 1
MLP	Window size: 12, hidden_layer_sizes: (50, 25), activation: 'relu', learning_rate: 'constant', learning_rate_init: 0.00045, max_iter: 120, tol: 0.0001, alpha: 0.0013, batch_size: 201, epsilon: 1e-05

The performance of classical and ML statistical models is shown in Figure 4. The metrics *MAE*, *MSE*, *RMSE*, *MAPE*, *SMAPE*, and *r* were calculated by evaluating the difference between the original rates and the rates predicted by the models, with the test dataset.

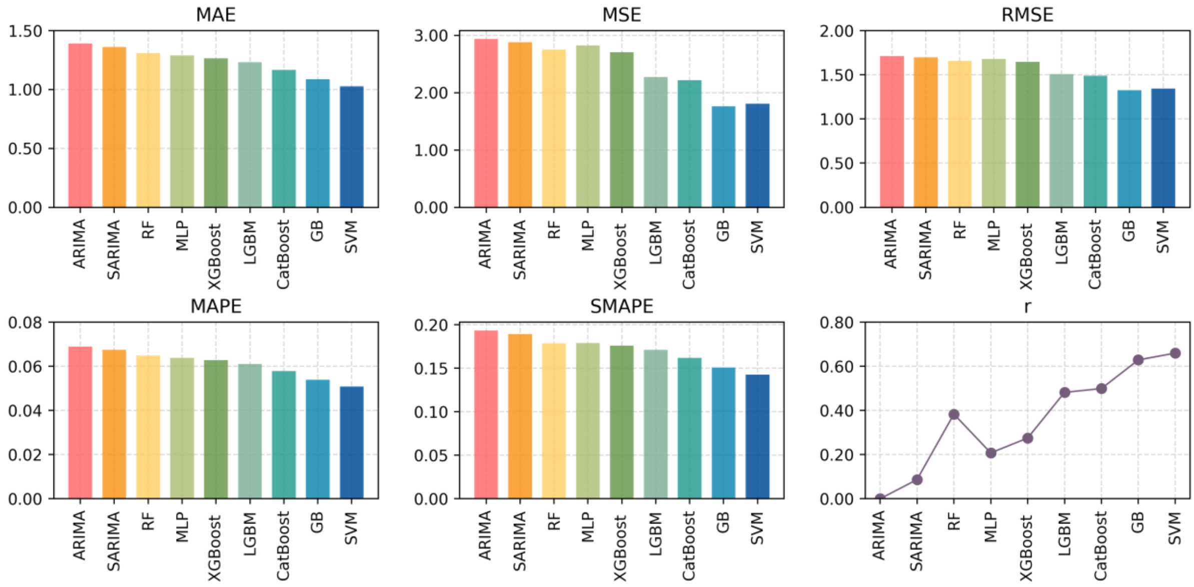

Figure 4: Performance metrics of classical and machine learning statistical models.

When comparing these models, the classic SARIMA (*MAE* = 1.362, *MSE* = 2.880, *RMSE* = 1.697, *MAPE* = 0.067, *SMAPE* = 0.189, *r* = 0.087) time series forecast model presented the highest values in the performance metrics, indicating low adequacy in the prediction of monthly neonatal mortality rates, as can be seen in Figure 4. However, when the performance of the SARIMA model was compared with the ML models, it became evident that the ML models tested had a better ability to predict monthly mortality rates.

From the performance of the ML models tested, the CatBoost (*MAE* = 1.168, *MSE* = 2.216, *RMSE* = 1.488, *MAPE* = 0.058, *SMAPE* = 0.162, r = 0.499), SVM (*MAE* = 1.027, *MSE* = 1.804, *RMSE* = 1.343, *MAPE* = 0.051, *SMAPE* = 0.143, r = 0.660), and GB (*MAE* = 1.088, *MSE* = 1.761, *RMSE* = 1.327, *MAPE* = 0.054, *SMAPE* = 0.151, *r* = 0.629) techniques presented the best evaluation metrics, that is, they presented the lowest values in the error metrics and the highest values in the correlation coefficient.

On the other hand, the LGBM (*MAE* = 1.232, *MSE* = 2.270, *RMSE* = 1.507, *MAPE* = 0.061, *SMAPE* = 0.171, *r* = 0.482), XGBoost (*MAE* = 1.266, *MSE* = 2.706, *RMSE* = 1.645, *MAPE* = 0.063, *SMAPE* = 0.176, *r* = 0.275), MLP (*MAE* = 1.289, *MSE* = 2.826, *RMSE* = 1.681, *MAPE* = 0.064, *SMAPE* = 0.179, *r* = 0.208), and RF (*MAE* = 1.310, *MSE* = 2.752, *RMSE* = 1.659, *MAPE* = 0.065, *SMAPE* = 0.179, *r* = 0.383) models presented the worst values in the evaluated metrics, but even so, they were lower values when compared to the SARIMA model.

Figure 5 illustrates the forecasts of monthly neonatal mortality rates, for the period from December/2018 to December/2020, based on the Catboost, SVM, and GB models, and the original rates.

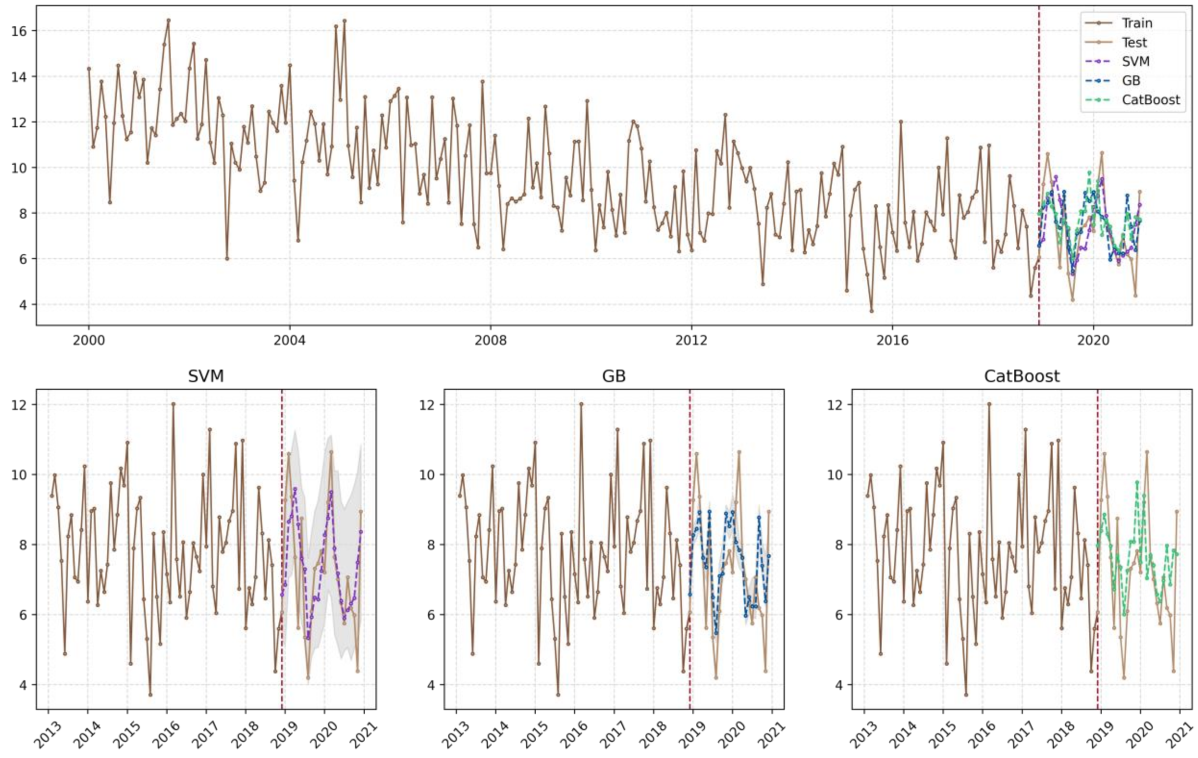

Figure 5: Original and predicted monthly neonatal mortality rates with the best machine learning models.

Through a visual analysis in Figure 5, the SVM model showed better behavior in forecasting monthly rates, showing a high degree of proximity between the original data and the forecasted values. Likewise, when evaluating its performance through error metrics, the SVM model presented the lowest values in error rates and the highest value in r. To confirm this result, a paired t-test was performed to compare the performance of the SVM (mean = 7.359, standard deviation = 1.193), CatBoost (mean = 7.671, standard deviation = 0.886), and GB (mean = 7.495, standard deviation = 1.057), as shown in Table 4. Note that the mean and standard deviation of the real data analyzed were 7,274 and 1,748, respectively.

Table 4
Results of the paired t-test between the SVM, CatBoost, and GB models.

Model	t-value	t-critical	p-value	Observation
SVM	0.311	2.064	0.759	Accept H_0
CatBoost	1.353	2.064	0.189	Accept H_0
GB	0.824	2.064	0.418	Accept H_0

The results of the paired t-test showed that the mean of the real data and the mean of the selected models were very close values, so the null hypothesis was accepted for the three models. Furthermore, comparing the t-value and t-critical results of the SVM, CatBoost, and GB models (Table 4), it was observed that the t-value of the SVM (0.311) model is lower than that of the GB (0.824), and CatBoost (1.353) models, indicating that the predictions with the SVM model is closer to the real values, unlike the other models. Thus, the prediction of events with time series related to neonatal deaths can significantly help managers in decision-making, such as resource management, strategic planning, development of public policies, improvement of care for newborns and mothers, etc., with the aim of mitigating risks and reducing the number of deaths [10, 12, 13].

4. Conclusion

Neonatal mortality is an important public health problem because newborns have a higher risk of death in the first 28 days of life. Worldwide, the mortality rate has decreased considerably and this decrease in deaths occurred mainly in the post-neonatal period. In Brazil, the same behavior has been observed when evaluating infant and neonatal mortality rates.

In this sense, the prediction of future events can significantly help in the reduction of deaths, as well as in the development of public policies to mitigate risks with preventive care for newborns and improve the assistance received.

In this work, classical models and ML models were compared to predict monthly neonatal mortality rates in the metropolitan region of Paraiba River Valley and North Coast – São Paulo State - Brazil, for the period from January 2000 to December 2020. The classical statistical model evaluated in this work was SARIMA, in addition to the supervised models of RF, SVM, LGBM, CatBoost, GB, XGBoost, and MLP.

The SARIMA model presented the highest values in the error metrics evaluated, compared to the ML models. The SVM model presented better precision, demonstrating that it was possible to capture the complex patterns of the data for the prediction of monthly neonatal mortality rates. The results were reinforced by the paired t-test, confirming that the SVM model predicted the mortality rates with a behavior very close to the actual rates.

Among the limitations of this work is the use of data provided by DATASUS, in addition to the comparison of ML models only with the SARIMA method. In future work, we intend to compare other classical time series models, as well as develop models for predicting mortality rates for the municipalities in the analyzed region.

Finally, it is important to highlight that prediction models are not a solution for neonatal mortality; however, they can significantly contribute to the approach of preventive measures for newborn care.

5. Acknowledgments

This work was supported by the Coordination for the Improvement of Higher Education Personnel [grant number CAPES - 001]; and partially by National Council for Scientific and Technological Development [grant numbers CNPq – 304197/2021-1; CNPq 303090/2021-9].

6. References

[1] L. Hug, M. Alexander, D. You, L. Alkema. "National, regional, and global levels and trends in neonatal mortality between 1990 and 2017, with scenario-based projections to 2030: a systematic analysis", Lancet Glob Health 7(6): e710–e720, 2019 . doi: 10.1016/S2214-109X(19)30163-9

[2] World Health Organisation. "Global Nutrition Targets 2025: Low birth weight policy brief", 2014. URL: https://apps.who.int/iris/bitstream/handle/10665/149020/WHO_NMH_NHD_14.5_ eng.pdf?ua=1

[3] World Health Organisation. "Newborns - improving survival and well-being", 2020. URL: https://www.who.int/news-room/fact-sheets/detail/newborns-reducing-mortality

[4] J.E. Lawn, H. Blencowe, S. Oza, D. You, A.C. Lee, P. Waiswa, et al. "Every Newborn: progress, priorities, and potential beyond survival", The lancet 384(9938): 189-205, 2014. doi: 10.1016/S0140-6736(14)60496-7

[5] United Nations. "The Millennium Development Goals Report 2015", 2016. URL: https://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20(July%201).pdf

[6] United Nations. "Transforming our world: the 2030 Agenda for Sustainable Development", 2015. URL: https://www.un.org/ga/search/view_doc.asp?symbol=A/%20RES/70/1&Lang=E

[7] C.D.S.R. Marinho, T.B.M. Flor, J.M.F. Pinheiro, M.Â.F. Ferreira. "Millennium Development Goals: the impact of healthcare interventions and changes in socioeconomic factors and sanitation on under-five mortality in Brazil", Cadernos de Saúde Pública 36(10):e00191219, 2020. doi: 10.1590/0102-311X00191219

[8] M.D.C. Leal, C.L. Szwarcwald, P.V.B. Almeida, E.M.L. Aquino, M.L. Barreto, et al. "Reproductive, maternal, neonatal and child health in the 30 years since the creation of the Unified Health System (SUS)", Ciência & Saúde Coletiva 23: 1915-1928, 2018. [Article in Portuguese, English] doi: 10.1590/1413-81232018236.03942018

[9] A.R.S. Parmezan, V.M. Souza, G.E. Batista, "Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model", Information sciences 484: 302-337, 2019. doi: 10.1016/j.ins.2019.01.076

[10] N. Shakhovska, I. Darmoriz, Y. Vyklyuk, Y. Kryvenchuk, P. Pukach, "Visualization of the Epidemics Forecasting Results". The 4th International Conference on Informatics and Data-Driven Medicine (IDDM 2021), 3038(2021): 283-292.

[11] F. Petropoulos, D. Apiletti, V. Assimakopoulos, M.Z. Babai, D.K. Barrow, S.B. Taieb, et al. "Forecasting: theory and practice", International Journal of Forecasting 38(3): 705-871, 2022). doi: 10.1016/j.ijforecast.2021.11.001

[12] K.J. Foreman, N. Marquez, A. Dolgert, K. Fukutaki, N. Fullman, M. McGaughey, et al. "Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016-40 for 195 countries and territories", Lancet 392(10159):2052-2090, 2018. doi: 10.1016/S0140-6736(18)31694-5

[13] I.N. Soyiri, D.D Reidpath. "An overview of health forecasting", Environ Health Prev Med 18(1):1-9, 2013. doi: 10.1007/s12199-012-0294-6

[14] A.B.S. Silva, A.C.M. Araújo, P.G. Frias, M.B.R Vilela, C.V.B. "Auto-Regressive Integrated Moving Average Model (ARIMA): conceptual and methodological aspects and applicability in infant mortality", Rev Bras Saúde Mater Infant 21(2):657-666, 2021. doi: 10.1590/1806-93042021000200016

[15] E.Y.A. Rodríguez, E.C.A. Rodríguez, A.F Silva, P.M.R. Rizol, R.C. Miranda, F.A.S. Marins (in press). "Analysis of Machine Learning integration into SupplyChain Management", Int J Logist Syst Manag, 2021.

[16] D.S. Char, N.H. Shah, D. Magnus. "Implementing Machine Learning in Health Care - Addressing Ethical Challenges", New England Journal of Medicine 378(11): 981-983, 2018. doi:10.1056/NEJMp1714229.

[17] E. Rodríguez, E. Rodríguez, L. Nascimento, A. Silva, F. Marins, "Machine learning techniques to predict overweight or obesity". The 4th International Conference on Informatics and Data-Driven Medicine (IDDM 2021), 3038(2021): 190-204.

[18] Information Technology Department of the Brazilian Unified Health System. "Ministry of Health: Health Information", 2022. [in Portuguese] URL: https://datasus.saude.gov.br/

[19] D. Kwiatkowski, P.C.B. Phillips, P. Schmidt, Y.Shin. "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?", Journal of Econometrics 54(1-3):159-178, 1992. doi: 10.1016/0304-4076(92)90104-Y

[20] W.A. Fuller, Introduction to Statistical Time Series, 1996. Canada: John Wiley & Sons, Inc.

[21] M. Cools, E. Moons, G. Wets, "Investigating the variability in daily traffic counts through use of ARIMAX and SARIMAX models: assessing the effect of holidays on two site locations", I. Transportation research record 2136(1):57-66, 2009.

[22] E.Y.A. Rodríguez, A.A.R. Gamboa, E.C.A. Rodríguez, A.F. da Silva, P.M.S.R. Rizol, F.A.S. Marins. "Comparison of adaptive neuro-fuzzy inference system (ANFIS) and machine learning algorithms for electricity production forecasting", IEEE Latin America Transactions 20(10):2288–2294, 2022.

[23] M. Tim, K. Ekrem, T. Burak, M. Leandro, P. Fayola. "Sharing Data and Models in Software Engineering", 2014.

[24] R. Collobert, S. Bengio. "SVMTorch: Support vector machines for large-scale regression problems", Journal of machine learning research, 1(Feb), 143-160, 2001.

[25] A. Galicia, R. Talavera-Llames, A. Troncoso, I. Koprinska, F. Martínez-Álvarez. "Multi-step forecasting for big data time series based on ensemble learning", Knowledge-Based Systems, 163:830-841, 2019.

[26] R.G. Makade, S. Chakrabarti, B. Jamil. "Real-time estimation and prediction of the mortality caused due to COVID-19 using particle swarm optimization and finding the most influential parameter", Infectious Disease Modelling 5:772-782, 2022. doi: 10.1016/j.idm.2020.09.003

[27] S.X. Lv, L. Peng, H. Hu, L. Wang. "Effective Machine Learning Model Combination based on Selective Ensemble Strategy for Time Series Forecasting", Information Sciences 612:994-1023, 2022.

[28] D. C. Montgomery and G. C. Runger, "Applied Statistics and Probability for Engineers". John Wiley and Sons, 6th ed., 2013.