Modeling in High School: Rate of Change Model for Predicting the Number of Cases and Deaths from Covid-19

Lima Pedro * and Pereira Valberto

Laboratory of Educational Data Analysis and Applied Statistics, Department of Physics and Mathematics, Federal Institute of Ceará. Fortaleza, Ceará, Brazil.

World Journal of Advanced Research and Reviews, 2024, 23(02), 1032–1043

Publication history: Received on 25 June 2024; revised on 04 August 2024; accepted on 06 August 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.23.2.2316

Abstract

This study presents an innovative approach to mathematical modeling in the context of high school education, introducing a new paradigm aimed at simplifying the understanding and application of this discipline for students in this age group. The core of this approach is the creation of the MTV model (Rate of Variation Model). This creation consistently demonstrated superior performance compared to traditional models such as the Gompertz model.

The novelty brought by MTV stands out by incorporating machine learning algorithms, which dynamically adapt the model parameters based on available data. This results in more accurate predictions and remarkable ability to adjust to variations in observed data.

Experiments conducted with different datasets reinforce the effectiveness of this new paradigm, demonstrating consistently superior performance compared to the Gompertz model. The flexibility shown in dealing with a variety of growth patterns and the ability to adapt to less structured data consolidate its robustness. Moreover, by allowing students to obtain results closer to reality, the approach promotes a deeper understanding of the underlying mathematical principles.

Consequently, the creation of the MTV Model, with its foundation in machine learning principles and dynamic adaptation, may lead to a new generation of high school students better prepared to tackle quantitative challenges in an ever-changing landscape.

Keywords: Mathematics; Statistics; Probability; Machine Learning; Gompertz.

1. Introduction

The recent pandemic, stemming from Covid-19, a disease-causing clinical presentation ranging from asymptomatic to severe cases, began in China, more precisely in a fish market in Wuhan, at the end of 2019 [1]. In the Brazilian context, the first case of Covid-19 emerged in São Paulo in February 2020 [2], after which we witnessed a relentless growth.

From these moments, the search for information increased sharply, especially in the healthcare sector [3], obviously in terms of mathematical prediction, considering that we had never experienced such a scenario on a global scale in modern times. From this point onwards, the need for work with statistical data became even more indispensable, however, these are derived from advanced mathematics, which complicates popular understanding.

Having said that, we observed that amidst this, access to and interpretation of data could be facilitated, especially with the use of software allied with some mathematical bases such as moving averages[1], models like the Gompertz law[2].

* Corresponding author: Lima Pedro

[1]Moving averages are popularly known among trend followers. They are widely used in the Day trading market.

[2]The Gompertz model is one of the oldest and most well-known models. This model is widely used, especially in biology and demographic studies. It has been frequently used to describe the growth of animals and plants, as well as to model human mortality. [4]

In addition to the issues raised, we also had some more pertinent ones, regarding how to deal with these numbers without all the mathematical formality and to present them in a clearer way. However, such methods often bring about certain confusions of ideas, considering that when dealing with more advanced mathematics, compliance with rules for non-ambiguity becomes almost mandatory.

During the context experienced during the research, at the height of the pandemic, the flow of information only served to raise further doubts regarding the disease and its current stage, while one press source claimed that the numbers showed that the pandemic was about to be controlled, another pointed out that we were on the brink of collapse.

Analogous to this, academic studies and research were extremely scarce on the specific subject, there was a study done on Ebola [5] that caught attention at the time, but perhaps it did not apply to the current scenario due to the geographical situation addressed, after all Ebola had a limited peak almost exclusively in the African continent and we were dealing with a global crisis.

Would it be possible to construct an effective model using less advanced calculations? And with what tools would we measure this accuracy?

The answer is yes, through the developed model, the MTV, it was possible to arrive at very close results, comparing the real values and the values estimated by the Gompertz model using the RMSE metric.

Before continuing the train of thought, it is interesting to discuss why these mathematical models are so widely used, even though they often do not present absolute results. To provide an explanation for this, we must explore the medical realm a bit, considering that the approaches necessary for the control of viruses, such as SARS-COV-2, must be based and structured for greater accuracy. From a prediction of cases, it is possible, for example, to know if the quantity of vaccines being produced now will be sufficient, that is, effective for the progression of virus combat and, consequently, the disease. [9]

2. Material and methods

2.1 Preliminary Explanations.

Let $C: N \to N$ be the function associating each day with the number of cumulative cases of Covid-19 infections. To solidify our understanding, let's consider Table 1.

Table 1 Government database in the state of Ceará, Brazil, (https://covid.saude.gov.br) Last accessed: August 7, 2020.

Data	Day	Accumulated cases
March 17	1	5
March 18	2	9
March 19	3	20
March 20	4	55
March 21	5	68
March 22	6	112
March 23	7	163
March 24	8	182
March 25	9	200

March 26	10	235
March 27	11	282
March 28	12	314

For a correct interpretation of the explanation below, take the notations above in the following way:

$$C(1) = 5, C(2) = 9\,, C(3) = 20, C(4) = 55, C(5) = 68.$$

The function (C) is unknown, which means we do not have access to its algebraic formula, and therefore, we cannot precisely determine the value of (Cn) for any value of n {N}, including n being a day in the future.

For example, if we are on March 21, 2020, corresponding to day 5, and we want to know the value of C (12), which would be March 28, 2020, it would not be possible to determine the exact value of C (12) without prior knowledge of the function.

However, it is possible to make estimates of the value of C (12) based on available data or information. These estimates can be made based on trends or patterns observed in historical data or other sources of information. It is important to note that these estimates are only approximations (thus, we always seek to improve the approximation, i.e., reduce the prediction error) and do not guarantee the accuracy of the value of C (12).

There is also another approach, more computational, with the aid of machine learning, there is a familiarity with the model operated in this way and the computational aspect. [8]

With that said, we will delve a little deeper into the mathematical concepts of our proposal.

2.1.1 Model Formalization (MTV)

Based on the observed data patterns, we constructed the function P, which we will now explain in terms of its domain and formulation.

Definition of function P:

$P: \sum \rightarrow N$ where $\sum = \{11k + 1, k \in N \; e \; k > 3\}$

$P(n_k) = C(n_{k-1}) \times MTC_{k-2,k-3} + C(n_{k-1})$

Where:

$P(n_k)$ is the estimate for the value of $C(n_k)$;

where n_k represents a day in the future, defined as $n_k = 11k + 1$, with $k \in N \; e \; k > 3$;

The average rate of change for $j \geq 1$ is given by:

$$MTC_{j,j-1} = \frac{TC(j) + TC(j-1)}{2}$$

The rate of change of (C) over ten days will be given by:

$$TC(\,j\,) = \frac{C(n_{j,11}) - C(n_{j,2})}{C(n_{j,2})} \times 100$$

The days in the past for calculating the rate of change will be given by:

$n_{j,i} = 11j + i, j \in \{0,1,2,3,..\}$ e $i \in \{2, 11\}$

Example

Let's assume we are on day $n_{k-1} = 11(k-1) + 1$, and we want to estimate the value of $C(n_k)$, using the function P, then we have:

$$C(n_k) \cong P(n_k) = C(n_{k-1}) \times MTC_{k-2,k-3} + C(n_{k-1})$$

Let's suppose we are on day $34 = 11 \times 3 + 1 = 11 \times (4-1) + 1$, which means $C(n_3) = C(34)$, so we can estimate $C(n_4)$ using $P(n_4) = P(45)$.

$P(n_4) = C(n_{4-1}) \times MTC_{4-2,4-3} + C(n_{4-1})$

$$P(n_4) = C(n_3) \times MTC_{2,1} + C(n_3)$$

$$MTC_{j,j-1} = \frac{TC(j) + TC(j-1)}{2} = MTC_{2,1} = \frac{TC(2) + TC(1)}{2}$$

Let's Calculate:

$$TC(\ j\ ) = \frac{C(n_{j,11}) - C(n_{j,2})}{C(n_{j,2})} \times 100$$

$$TC(\ 2\ ) = \frac{C(n_{2,11}) - C(n_{2,2})}{C(n_{2,2})} \times 100$$

$$TC(\ 1\ ) = \frac{C(n_{1,11}) - C(n_{1,2})}{C(n_{1,2})} \times 100$$

Remembering that:

$$n_{j,i} = 11j + i$$

Hence:

$$C(n_{2,11}) = C(11 \times 2 + 11) = C(33) = 3.024\ Cumulative\ Cases$$

$$C(n_{2,2}) = C(11 \times 2 + 2) = C(24) = 1.425\ Cumulative\ Cases$$

$$TC(\ 2\ ) = \frac{3.034 - 1.425}{1.425} \times 100 = 112{,}91\%$$

$$C(n_{1,11}) = C(11 \times 1 + 11) = C(22) = 1.051\ Cumulative\ Cases$$

$$C(n_{1,2}) = C(11 \times 1 + 2) = C(13) = 348\ Cumulative\ Cases$$

$$\mathrm{M}TC(\ 1\ ) = \frac{1.051 - 348}{348} \times 100 = 202{,}01\%$$

Rate of Change:

$$MTC_{2,1} = \frac{TC(2) + TC(1)}{2} = \frac{112{,}91 + 202{,}01}{2} = 157{,}46\%$$

Estimation:

$$P(n_4) = C(n_3) \times MTC_{2,1} + C(n_3) = C(34) \times MTC_{2,1} + C(34)$$

$$P(n_4) = P(45) = 3.024 \times 157{,}46\% + 3.024 = 7.811$$

Day 45 corresponds to April 30, 2020. On this day, C45 = 7,606.

2.2 Comparison Metric (ANNEX A) [22]

To assess the accuracy of our predictions, we used the Root Mean Square Error (RMSE) metric to obtain data regarding the accuracy of our model.

RMSE is expressed in the same units as the original values, which facilitates interpretation. The lower the RMSE value, the closer the model's predictions are to the actual values, indicating greater model accuracy. RMSE is especially useful when errors have a significant effect on the problem at hand and when evaluating the magnitude of prediction errors.

The metric provides us with more precise information regarding the estimates. For example, a study on the accuracy of estimates made for precipitation in a certain area using satellite data showed that despite the medium being used as a prediction tool, the metric indicated it to be less efficient [10].

The RMSE metric is frequently used to evaluate models in physical sciences such as meteorology and physical oceanography. For instance, Atwater and Ball [17] utilize it. Additionally, Dutta [18] evaluates the performance of three neural network models for COVID prediction using RMSE. Gachoki [19] uses RMSE to assess models of plant growth, logistic regression models, and Gompertz models used to predict peaks of confirmed cases and deaths in Cuba, as well as the total number of cases, have their performances evaluated by RMSE [20].

RMSE is calculated as follows: [21]

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2}$$

O_i = (Observed Case Numbers): Represents the actual number of COVID-19 cases recorded in each period.

P_i = (Predicted Case Numbers): Denotes the number of COVID-19 cases forecasted by the mathematical model for the same period.

To evaluate the application in the context of this research, please refer to ANNEX A. [22]

2.3 Precision. (Gompertz Model and its Calibration)

As seen previously, it is possible to forecast cases by obtaining real data to establish a sort of baseline. Using this real case baseline, we see the possibility of predicting the emergence of new cases of Covid-19. Consequently, every 10 days, the accuracy of the cases is tested and, if necessary, corrected or redone. By comparing the actual and predicted numbers, it was already possible to notice a divergence of values when the forecast was misaligned. To confirm this hypothesis, the Curve Expert Basic software [7] can be used. Upon discovering an error, it was necessary to restart the calculations from day one, thus modifying the constants of the Gompertz curve.

We adopted the Gompertz model to measure the performance of the MTV model using the RMSE metric. The Gompertz model is one of the oldest population growth models, with its use dating back to 1825 by Benjamin Gompertz [11]. Its application is diverse, being used to predict the growth of plants, bacteria, and mortality [12]. It has been widely used in the recent pandemic to verify model performance or to work in conjunction with other models in Covid predictions [13-16].

The Gompertz differential equation is shown as follows:

$$\frac{dy}{dt} = cyln\left(\frac{1}{ky}\right), c\ e\ k\ are\ positive\ constants, \text{and } t\ is\ time \qquad \text{I}$$

By defining $a = \frac{1}{k}$, we can rewrite the equation as:

$$\frac{dy}{dt} = cyln\left(\frac{a}{y}\right) \qquad \text{II}$$

To solve this equation using the method of separation of variables, we rearrange the terms:

$$\frac{dy}{yln(\frac{a}{y})} = cdt \qquad \text{III}$$

Since $ln(\frac{a}{y}) = lna - lny$, substituting this into equation III yields:

$$\frac{dy}{y(lna - lny)} = cdt \qquad \text{IV}$$

Integrating both sides, we have:

$$\int \frac{dy}{y(lna - lny)} = \int \text{cdt} \quad \text{V}$$

Letting, we find the differential $\frac{du}{dy}$:

$$\frac{du}{dy} = -\frac{1}{y} \qquad \text{VI}$$

$$dy = -ydu \quad \text{VII}$$

Substituting $dy = y$ into equation (V), we obtain:

$$\int \frac{du}{u} = -\int \text{cdt}$$

$$lnu = -ct + b \text{ , } where\ b\ is\ a\ constant$$

$$u = e^{-ct+b}$$

$$lna - lny = e^{-ct+b}$$

$$ln\, ln \left(\frac{a}{y}\right) = e^{-ct+b}$$

$$a/y = e^{e^{-ct+b}}$$

$$y/a = e^{-e^{-ct+b}}$$

$$y(t) = ae^{-e^{(b-ct)}} \qquad \text{VIII}$$

Equation VIII will be employed to forecast the number of COVID-19 cases. To achieve accurate predictions, it is necessary to estimate the parameters a, b, and c. This can be accomplished using the software Curve Expert Basic (Hyams Development), which is used to calibrate the model.

During the calibration process, the software utilizes a dataset to estimate the values of the parameters a, b, and c, and provides the coefficient of determination (r^2) to evaluate the fit of the model. Once the model is calibrated, we proceed with making predictions. If the predictions diverge significantly from the observed data, we recalibrate the model using updated data.

2.4 Softwares used.

Microsoft Excel was used for data management, calculations, and predictions, while the Curve Expert Basic software (Hyams Development) was utilized for adjustments, calibrations, and calculation of parameters for the Gompertz model for precise monitoring. [6-7]

2.5 Regarding Data Acquisition and Monitoring

Data monitoring involved daily updates of real cases obtained through the Brazilian data disclosure portal, which can be accessed in the official Brazilian government database [2]. Data collection occurred from March 17 to August 7, 2020, and corresponds only to cases in the state of Ceará, Brazil.

3. Results and discussion

3.1 Adopting the Gompertz Model

The utilization of the Gompertz model allowed for achieving accuracies remarkably close to the real ones. However, it was necessary to recalibrate, that is, realign the accuracies several times, considering that if we were to use only the first sampling of the model, the error would be significant, given that the numbers do not grow in an ordered manner and do not follow specific patterns. Therefore, a period of analysis (A.P.) was needed, representing a period necessary to accumulate the cases that occurred during that period and, from there, calibrate the model for use in the forecast. It is noteworthy that there were 10 calibrations in total, but for analysis purposes, only some of them will be used. For an overview of the entire table, refer to ANNEX B [23]. See **Table 2.**

Table 2 Forecasts using the Gompertz model.

Data	Accumulated cases	Gompertz Model 1	Gompertz Model 3	Gompertz Model 5	Gompertz Model 7
Mar 17	5	A.P.	A.P.	A.P.	A.P.
Mar 28	314	A.P.	A.P.	A.P.	A.P.
Apr 8	1.291	A.P.	A.P.	A.P.	A.P.
Apr 19	3.252	A.P.	A.P.	A.P.	A.P.
Apr 30	7.606	6.827	A.P.	A.P.	A.P.
May 11	17.599	12.117	17.765	A.P.	A.P.
May 22	34.573	18.649	37.939	A.P.	A.P.
Jun 2	53.073	25.785	75.458	50.271	A.P.
Jun 13	76.429	32.894	140.703	71.575	A.P
Jun 23	99.578	39.499	247.468	93.958	100.665
Jul 5	121.464	45.322	412.791	115.861	124.834
Jul 16	144.000	50.257	656.292	136.149	147.183
Jul 27	162.429	54.315	999.018	154.164	166.954
Aug 7	185.409	57.580	1.461.948	169.646	183.861
Aug 18	199.258	60.162	2.064.327	182.619	197.948
Aug 29	214.094	62.178	2.822.072	193.283	209.454

In Table 2, some forecasts adopted using the Gompertz model are represented. It is interesting to note that with their recalibration, that is, the adjustment made initially, the forecasts become more accurate. This is evident in each of them; for instance, in the first and second calibrations, the error is significant, but with the readjustment, forecasts 3 and 4 become closer to the real numbers. Thus, the more calibrations are performed, the more accurate the result will be.

3.2 Adopting the MTV Model

With another approach, we have the MTV Model, developed and applied by LADE, which can bring us even closer to the real numbers. Refer to Table 3 for details.

Table 3 Forecasts using the MTV model

Data	Accumulated cases	MTV Model
Mar 17	5	A.P.
Mar 28	314	A.P.
Apr 8	1.291	A.P.
Apr 19	3.252	A.P.
Apr 30	7.606	7.811
May 11	17.599	15.319
May 22	34.573	35.100
Jun 2	53.073	66.115
Jun 13	76.429	83.433
Jun 23	99.578	106.561
Jul 5	121.464	130.511
Jul 16	144.000	149.103
Jul 27	162.429	168.307
Aug 7	185.409	183.862
Aug 18	199.258	205.606
Aug 29	214.094	215.211

With the adoption of the MTV model, the extensive calibration and adjustment processes that were previously mandatory become unnecessary. This is due to the intrinsic ability of the model to adjust automatically, making it an incredibly more accessible method, requiring considerably fewer checks of its accuracy.

3.3 Comparison of the Models Addressed Using the RMSE Metric.

Table 4 Error descriptions and their means.

Gompertz Models	Gompertz-RMSE	MTV-RMSE	Calibration
1	70.106	6.467	17/03 - 19/04
2	150.515	6.467	17/03 - 29/04
3	548.053	6.817	17/03 - 10/05
4	143.237	7.185	17/03 - 21/05
5	8.219	7.679	17/03 - 01/06
6	39.723	7.679	17/03 - 12/06
7	3.135	6.222	17/03 - 23/06
8	5.981	6.017	17/03 - 04/07
9	4.795	6.017	17/03 - 15/07
10	5.913	4.298	17/03 - 26/07

Based on the previous sections, we can conclude that the MTV model demonstrated greater accuracy in its final forecasts when compared to the actual values. This assertion is based on the final results presented in Table 2, corresponding to the 7th forecast, as well as Table 3, which corresponds to August 29th. However, when we aggregate all the measurements conducted, a different perspective emerges, as presented in Table 4.

As we have seen previously, the MTV model is easier to apply than the Gompertz model, thus reinforcing our main objective of facilitating these predictions. To observe this point more clearly, please refer to ANNEX B. [23]

In this way, we can apply it and obtain results that require less mathematical effort, aiming at our main goal, which is to streamline this process. Thus, we have not only eliminated the need for much more advanced mathematics but also achieved relatively more accurate precision.

Table 4 shows the comparative errors between the two metrics. Note that only in 3 cases did the Gompertz model perform better, but in most cases, the MTV model played a more precise role, thus being more efficient and less laborious.

4. Conclusion

In summary, the present study outlined an innovative approach to mathematical modeling in the context of high school education, through the introduction of a new paradigm. This approach, grounded in the concept of exponential learning, aims to simplify the understanding and application of mathematical modeling for students in this age group. The results obtained demonstrated consistent superiority of our model (MTV) over traditional models, such as the Gompertz model.

The proposed method incorporates machine learning algorithms, which dynamically adjust model parameters based on available data. This results in more accurate predictions and flexible adaptation to variations in observed data.

The practical implementation of this new approach on different datasets has shown consistently superior performance compared to the Gompertz model. Its flexibility to handle various growth patterns and its ability to adapt to less structured data strengthen its effectiveness. Furthermore, by allowing students to obtain results closer to reality, the new paradigm promotes a deeper understanding of the underlying mathematical fundamentals.

Ultimately, this innovative approach in mathematical modeling in high school education, based on machine learning principles and dynamic adjustment, can lead to a new generation of more proficient students ready to tackle quantitative challenges in an ever-evolving world.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Worobey M, Levy JI, Malpica Serrano L, et al. The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic. Science. 2022;377(6609):951-959. doi:10.1126/science.abp8715

[2] Ministério da Saúde do Brasil. Coronavírus Brasil [Internet]. Brasília: Ministério da Saúde; 2021 [cited 2020 Aug 7]. Available from: https://covid.saude.gov.br/.

[3] Roberto FL, Silva P, Santos MA, et al. Information seeking on COVID-19 on the web: a cybermetric perspective. Rev Eletr Comun Inf Inov Saude. 2022;16(1). [Published in Portuguese]

[4] Castellares F, Patricio SC, Lemonte AJ. On gamma-Gompertz life expectancy. Stat Probab Lett. 2020;165:108832. doi:10.1016/j.spl.2020.108832.

[5] Bürger R, Chowell G, Lara-Díaz LY. Comparative analysis of phenomenological growth models applied to epidemic outbreaks. Math Biosci Eng. 2019;16(5):4250-4273. doi:10.3934/mbe.2019212.

[6] Microsoft Corporation. Microsoft Excel [software]. Redmond, WA: Microsoft Corporation; 2021.

[7] CurveExpert [computer program]. Version 2.7.5. Hyams Development, 2021.

[8] Batista EDA, Araújo WC, Lira RV, Batista LIA. Predicting dengue cases through Machine Learning and Deep Learning: a systematic review. Res. Soc. Dev. [Internet]. 2021 Nov; 10(11):e33101119347. Available from: https://doi.org/10.33448/rsd-v10i11.19347

[9] Barrozo S. A mathematical model for studying the dynamics and control of schistosomiasis transmission considering concomitant immunity and vaccination [dissertation]. Rio de Janeiro: Fundação Oswaldo Cruz; 2000. [Published in Portuguese].

[10] Viana DR, Ferreira NJ, Conforte JC. Evaluation of TRMM satellite 3B42 and 3B43 precipitation estimates in southern Brazil. In: Congresso Brasileiro de Meteorologia; 2010. [Published in Portuguese].

[11] Gompertz B. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 1825;182:513–85.

[12] TJØRVE, Kathleen MC; TJØRVE, Even. The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the Unified-Richards family. PloS one, v. 12, n. 6, p. e0178691, 2017.

[13] CONDE-GUTIÉRREZ, R. A.; COLORADO, D.; HERNÁNDEZ-BAUTISTA, S. L. Comparison of an artificial neural network and Gompertz model for predicting the dynamics of deaths from COVID-19 in México. Nonlinear Dynamics, v. 104, n. 4, p. 4655-4669, 2021.

[14] PELINOVSKY, E. et al. Gompertz model in COVID-19 spreading simulation. Chaos, Solitons & Fractals, v. 154, p. 111699, 2022.

[15] TORREALBA-RODRIGUEZ, O.; CONDE-GUTIÉRREZ, R. A.; HERNÁNDEZ-JAVIER, A. L. Modeling and prediction of COVID-19 in Mexico applying mathematical and computational models. Chaos, Solitons & Fractals, v. 138, p. 109946, 2020.

[16] AHMADI, Ali et al. Modeling and forecasting trend of COVID-19 epidemic in Iran until May 13, 2020. Medical Journal of the Islamic Republic of Iran, v. 34, p. 27, 2020.

[17] Atwater, M. A. and Ball, J. T. (1978) A numerical solar radiation model based on standard meteorological observations. Sol. Energy, Vol. 21,163-170.

[18] DUTTA, Shawni; BANDYOPADHYAY, Samir Kumar. Machine learning approach for confirmation of covid-19 cases: Positive, negative, death and release. MedRxiv, p. 2020.03. 25.20043505, 2020.

[19] GACHOKI, Peter; MURAYA, Moses; NJOROGE, Gladys. Modelling Plant Growth Based on Gompertz, Logistic Curve, Extreme Gradient Boosting and Light Gradient Boosting Models Using High Dimensional Image Derived Maize (Zea mays L.) Phenomic Data. American Journal of Applied Mathematics and Statistics, v. 10, n. 2, p. 52-64, 2022.

[20] MEDINA-MENDIETA, Juan Felipe; CORTÉS-CORTÉS, Manuel; CORTÉS-IGLESIAS, Manuel. COVID-19 forecasts for Cuba using logistic regression and gompertz curves. MEDICC review, v. 22, p. 32-39, 2022.

[21] GÉRON, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.", 2022.

[22] Attachment A: Applying the RMSE metric.

[23] Attachment B: Predictions on Microsoft Excel.

Appendix

Attachment A

Targeting the topic, that is; to facilitate understanding, we will address the practical application of the RMSE metric. For this, we will use Microsoft Excel.

Before we begin, let's look at the equation for the RMSE metric.

$$RMSE= E^2n$$

1st Step. We need a database and a prediction of cases.

To approach it in a more didactic way, we will use randomly generated numbers. As shown in the table:

Case Number	Forecast
5	4
14	12
22	19
48	50
52	53
79	85
81	80
180	201
286	220
364	402
392	380

2nd Step. Now let's calculate the error, in a simple way.

In the column next to the data of the number of cases and predictions, create a table for errors.

In the second row of the column, which refers to the first obtained data, use a subtraction equation by subtracting the value of real cases from the predicted ones.

Look at the example:

Case Number	Forecast	error
5	4	=B3-C3

After that, apply the step for each row of the table.

3rd Step. Now let's start making the calculations, starting with the calculation of our (E^2).

In Excel in Portuguese, we will use the function "=SOMAQUAD()".

In the English version, use the function "=SUMSQ()".

Within the parentheses, you should input the ranges of errors.

Here's an example:

Case Number	Forecast	error	Step Three
5	4	1	=SOMAQUAD(D3:D13)
14	12	2	
22	19	3	
48	50	-2	
52	53	-1	
79	85	-6	
81	80	1	
180	201	-21	
286	220	66	
364	402	-38	
392	380	12	

If done correctly, the result should be 6441.

4th Step. Now let's find our **N** for the equation, which is simply the number of cases and predictions.

In Excel in Portuguese, we will use the function "=CONT.VALORES()".

In the English version, use the function "=COUNT()".

Within the parentheses, you should input the ranges of errors, predictions, or even cases. In this step, we are not interested in the values themselves, but rather the quantity of data we have.

Case Number	Forecast	error
5	4	1
14	12	2
22	19	3
48	50	-2
52	53	-1
79	85	-6
81	80	1
180	201	-21
286	220	66
364	402	-38
392	380	12

Step Four

=CONT.VALORES(D3:D13)

If done correctly, the result should be 11.

5th Step. Concluding our RMSE calculation, now that we have all the necessary terms, let's apply the formula.

In Excel in Portuguese, we will use the function "=RAIZ()".

In the English version, use the function "=SQRT()".

Within the parentheses, you should input the results obtained in step 3 and divide it by the result of step 4.

In other words, you will get the square root of the division of E^2/N.

Look at the example:

Step Five
6441
11
=RAIZ(F3/F4)

If the steps were executed correctly, you should obtain the result **24.1980465.**

Attachment B

For full access to the spreadsheet containing all the data and records for this scientific article, the following link will provide the complete content in an Excel document.

https://docs.google.com/spreadsheets/d/1V3DLUzB7OSwG-PoxQXUw8DCekvLA93NV/edit?usp=sharing&ouid=106249023271709224278&rtpof=true&sd=true