INVESTIGACIÓN e INNOVACIÓN en INGENIERÍAS
ISSN2344-8652

Application of polynomial regression for the characterization of the COVID-19 curve using machine learning techniques

Aplicación de la regresión polinomial para la caracterización de la curva del COVID-19, mediante técnicasde machine learning

Gabriel Elías Chanchí Golondrino
University of Cartagena, Colombia

Wilmar Yesid Campo Muñoz
University of Quindío, Colombia

Luz Marina Sierra Martínez
University of Cauca, Colombia

Open Access

Received:
April 21, 2020

Accepted:
July 3, 2020

Published:
July 31, 2020

Correspondence:
gchanchig@unicartagena.edu.co

DOI:
https://doi.org/10.17081/invinno.8.2.4103

Abstract

Objective: This study proposes the application of polynomial regression for characterizing behavior patterns of COVID-19 cases, deaths, and recoveries in Colombia.

Methodology: The datasets were obtained from the reports provided by the Colombian Ministry of Health. First, the authors assessed the procedure used for performing polynomial regression through multiple linear regression. Then, a tool was designed and implemented for applying this procedure to the collected datasets. Finally, the corresponding results were assessed.

Results: For the three studies considered (cases, deaths, and recoveries), 20 polynomials were obtained along with the root mean square error and determination level associated with each study. Moreover, a set of predictions was generated based on the linear regressions obtained for each study.

Conclusions: Because polynomials with future values are volatile, polynomial equations have proven more useful when assessing COVID-19 curves up to the data capture date. This means that they can be used to determine the impact of government measures over a given period of time. Moreover, the machine learning libraries provided in Python considerably supported the estimation of polynomial regressions.

Keywords: COVID-19, machine learning, multiple linear regression, polynomial regression

Resumen

Objetivo: Caracterizar el comportamiento de las curvas de casos, muertes y personas recuperadas por el COVID-19 en Colombia, en esta investigación se propuso como aporte el uso del enfoque de regresión polinomial para el modelar el comportamiento de los datos.

Metodología: Se obtuvieron los datos a partir de los reportes proporcionados por el Ministerio de Salud de Colombia. En primera instancia, se estudió el procedimiento empleado para la obtención de la regresión polinomial mediante la regresión lineal múltiple, en segunda instancia, se diseñó e implementó una herramienta para la aplicación del procedimiento estudiado sobre el conjunto de datos recolectado, y finalmente, se realizó el análisis de los resultados.

Resultados: Se obtuvo para los tres estudios considerados (casos, muertes y personas recuperadas) 20 polinomios en conjunto con el error medio cuadrático (RMSE) y el nivel de determinación (R2) asociados a cada uno. Así mismo, se obtuvo un conjunto de predicciones basadas en las regresiones lineales obtenidas para cada estudio.

Conclusiones: La volatilidad de los polinomios con valores futuros, las ecuaciones polinómicas son más útiles para evaluar el comportamiento de la curva del COVID-19 hasta el día de captura de los datos, it means, pueden ser usadas para determinar el impacto de las medidas gubernamentales en un periodo determinado de tiempo. Igualmente, las librerías de machine learning provistas por el lenguaje Python, demostraron ser de gran apoyo a la estimación de la regresión polinomial.

Palabras clave: COVID-19, machine learning, regresión lineal múltiple, regresión polinomial.

© Copyright: Research and Innovation in Engineering

How to Cite (IEEE): G. E. Chanchí - Golondrino., W. Y. Campo - Muñoz., y L. M. Sierra - Martínez, "Application of polynomial regression for the characterization of the COVID-19 curve, using machine learning techniques," Revista Investigación e Innovación en Ingenierías, vol. 8, no. 2, pp. 87–105, 2020, https://doi.org/10.17081/invinno.8.2.4103

Introduction

Coronaviruses (CoVs) are viruses that appear periodically in different parts of the world and cause the so-called acute respiratory infection with varying levels of severity. COVID-19 is currently being transmitted on a large scale among people around the world. The World Health Organization reported the occurrence of different cases of severe acute respiratory infection caused by COVID-19 in Wuhan (China) since the last week of December 2019. Although the possible intensity of transmission is still unknown, the virus spreads when a person sneezes or coughs and expels droplets containing the virus, which come into contact with people nearby [1, 2, 3, 4]. The COVID-19 pandemic has been one of the most complex challenges that humankind has faced in recent history. The overall damage to human lives and its impact on different sectors of the economy remains unknown. The degree of contagion has grown to such an extent that it has caused a crisis in the health systems of different countries, in addition to the economic collapse that will severely affect the well-being of the population [3, 4, 5]. Particularly within the Colombian region, the first case of COVID-19 infection occurred on March 6, 2020, and the first two deaths occurred on March 22, 2020. The Ministry of Health has been issuing information daily on different means of communication regarding the number of cases, the number of deaths, and the number of people recovering from COVID-19. These reports update the information according to the tests that are being conducted but do not include studies on the characterization of data behavior using polynomial regressions.

Polynomial regression can be considered a particular case of multiple linear regression, in which one seeks to determine the best polynomial representing the data of a set of points [6, 7, 8, 9, 10]. In Colombia, obtaining a set of polynomials that estimate the growing trend of the cases, deaths and recovered people curves can be of great interest for decision-making by government entities. Machine learning supports users in conducting research, identifying patterns, detecting anomalies, and advocating new procedures [11, 12, 13, 14, 15]. Moreover, due to the current machine learning tools, it is possible to apply certain algorithms on a dataset, such as the linear regression algorithm, to obtain a prediction about the future behavior of the data [16, 17, 18, 19].

https://doi.org/10.17081/invinno.8.2.4103

Although there are tools for obtaining trends associated with polynomial regression, such as Excel, they have a limitation: the estimation of the automatic polynomial trend only addresses up to 6th-degree polynomials. Moreover, it is impossible to directly obtain the value of the root mean square error (RMSE). Hence, in this study, we used machine learning tools provided by the Python language, such as the scikit-learn library [17, 18], to personalize and extend the study of polynomial regression up to 20th-degree polynomials on data related to COVID-19 in Colombia.

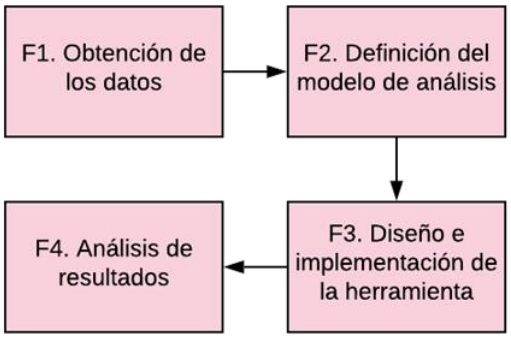
Hence, this study demonstrates the development of a study based on the use of the polynomial regression machine learning algorithm to determine the growth behavior of the COVID-19 cases, deaths, and people recovered curves in Colombia. As a support for the development of this study, a tool, which uses the machine learning libraries scikit-learn, numpy and matplotlib, were developed in the Python language. This tool analyzes data and obtains the polynomial regression equation for polynomials of different degrees [19, 20, 21]. The data analyzed in this study were obtained from the reports made by the Colombian Ministry of Health.

The study reveals how the trend curves based on polynomial regression follow the data more accurately than other types of mathematical models; however, because polynomial equations have a great variation in the long term, the greatest use of polynomial regression can be found in the assessment of the measures taken during the quarantine period for future decision-making. Moreover, this study aims to serve as a basis for extrapolation to other application contexts.

Methodology

For the development of this study, the following methodological phases were considered: obtaining the data, defining the analysis model, designing and implementing the tool, and analyzing the results, as shown in Figure 1.

Figure 1. Proposed methodology



Source: Prepared by the authors

In the first phase, data on cases, deaths, and people recovered from COVID-19 in Colombia were obtained from the official site and communication channels on the social networks of the Colombian Ministry of Health. The data contained records since March 6, 2020, when the first case of contagion was confirmed in Colombia. In the second phase, based on an initial analysis of the data and its comparison with other models (exponential and logarithmic), we decided to perform a polynomial regression study on the data obtained in phase 1 through multiple linear regression. In phase 3, based on the model definition, a tool in the Python language, which uses the machine learning libraries scikit-learn, numpy, and matplotlib, was designed and implemented to conduct the study. Finally, in phase 4, based on the designed tool, the polynomial regression study was applied to the data associated with the identified cases, deaths, and recovered people using polynomials of the 1st to the 20th degree.

Polynomial regression

Polynomial regression is a form of multiple linear regression in which one seeks to obtain the prediction of a quantitative response variable from a quantitative predictor variable, where the relation is modeled as a polynomial function of order or degree [22, 23, 24]. Through polynomial regression, it is possible to obtain a polynomial as the one shown in Equation (1), where ε represents the error in the estimation or the difference between the estimated value and the observed value.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots + \beta_p X^p + \varepsilon \quad . \tag{1}$$

The polynomial in Equation (1) can be expressed in the linear model of Equation (2), which can be approached as already mentioned, through multiple linear regression, where X1 = X, X2 = X2,...., Xp = Xp.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p + \varepsilon \quad . \tag{2}$$

To obtain multiple linear regression, it is possible to propose the matrix model for n data samples as presented in Equation (3).

$$\begin{pmatrix} y1 \\ y2 \\ . \\ . \\ yn \end{pmatrix} = \begin{pmatrix} 1 & x11 & \dots & x1p \\ 1 & x21 & \dots & x2p \\ 1 & . & \dots & . \\ 1 & . & \dots & . \\ 1 & xn1 & \dots & xnp \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ . \\ . \\ \epsilon_p \end{pmatrix} \quad . \tag{3}$$

Hence, using matrix notation, it is possible to express Equation (3) as shown in Equation (4).

$$Y = X\beta + \varepsilon \quad . \tag{4}$$

As the objective of Equation (4) is to find the vector of β coefficients, it is possible by matrix properties (transpose matrix and inverse matrix) to express Equation (4) in terms of Equation (5).

$$\beta = (X^T X)^{-1} X^T Y \quad . \tag{5}$$

Hence, the resulting vector, β, contains the different (β_0, β_1, β_2, …, β_p) coefficients of the polynomial presented in Equation (1). In this study, the mentioned mathematical procedure was used through the functionalities provided by Python's scikit-learn library, which approaches the problem as a special case of linear regression.

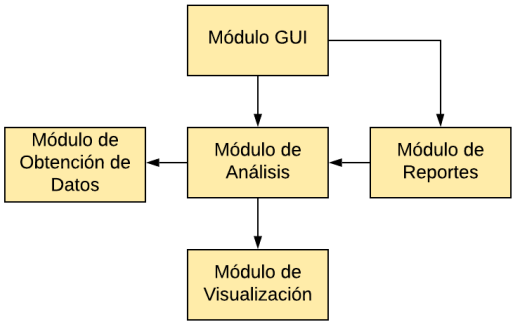
Proposed tool

As a contribution, this study proposes a tool to determine the growth behavior of the curves of cases, deaths, and people recovered from COVID-19 using polynomial regression algorithms and machine learning tools. Hence, this section presents the functional structure and implementation of the proposed tool.

Figure 2 presents the different functional modules the proposed tool is comprised of, among which we can find the graphical user interface, data acquisition, analysis, visualization, and reports modules [25, 26, 27].

The GUI module is responsible for managing the different graphical components of the interface (buttons, text boxes, and graphical components) as well as managing the different events that involve the user's interaction with said components. In the case of the proposed tool, this module was implemented from the Python Tkinter library, which allowed the construction of the graphical interface and the articulation of the different interaction components. Then, the different options to be considered for the analysis are chosen in the tool's GUI (study type and polynomial degree to be considered).

Figure 2. Functional modules



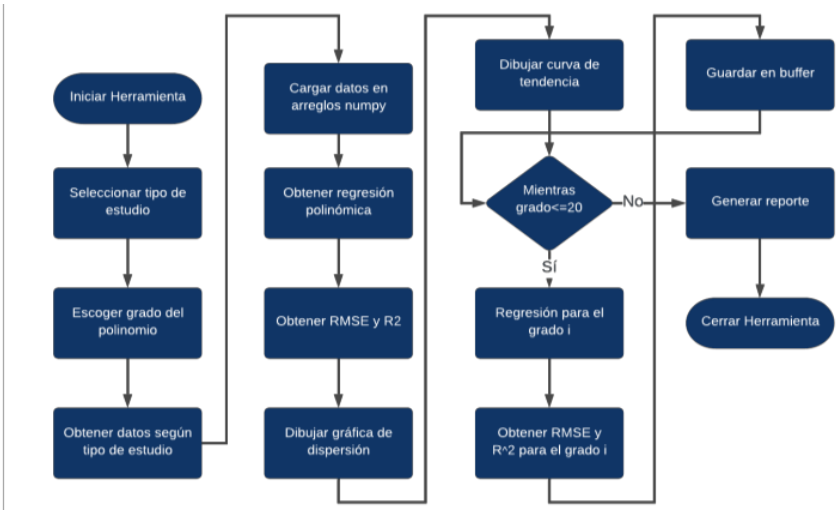
Source: Prepared by the authors

The analysis module obtains and analyzes the historical data of cases, deaths, and people recovered from COVID-19 using the polynomial regression algorithm, which considers the polynomial degree and the study type (cases, deaths, and recovered people) to be conducted as chosen by the user. The data processed by the analysis module is obtained with the support of the data acquisition module, which loads the data from plain text files and transforms them into arrays of the Python numpy library. To obtain the polynomial regression, the analysis module uses the functionalities provided by the Python scikit-learn machine learning library. Once the data are analyzed, the RMSE and determination level (R^2) of the trend curve are assessed regarding the obtained data (closeness of the trend to the data) and, with the visualization module, a graph is generated with the original points and the obtained trend curve, a process that is performed using the Python matplotlib library. Finally, with the report module, the user can generate a comparative report with the different trends based on the study type (cases, deaths, and recovered people) by considering the different degrees of polynomial considered in the tool and the error of the trend line related to the data associated with each degree.

https://doi.org/10.17081/invinno.8.2.4103

Figure 3 contains a flowchart that illustrates the process followed by the tool to perform the analysis of the COVID-19 data and obtain the trend curves. In the first instance, once the tool is started, the user chooses the study type to be conducted (cases, deaths, and recovered people) and the polynomial degree to be assessed, i.e., the polynomial degree that will be used to determine the data trend. From the above, the tool is responsible for obtaining the data set considering the study type and loading them into arrays of the Python numpy library. Then, the tool is responsible for applying polynomial regression on the data depending on the type of polynomial degree that the user has chosen, thereby obtaining an equation that represents the data behavior. Additionally, the tool calculates the trend equation, RMSE, and R^2, which show how close the trend is to the analyzed data. Both the regression and its assessment are performed based on the functionalities provided by the Python scikit-learn machine learning library. Next, using the Python matplotlib library, a scatter plot is generated with the loaded data and the trend curve representing the equation obtained for the degree of the chosen polynomial. Finally, the user can generate a report of the selected study type, for which the tool calculates the RMSE and R^2 on the loaded data for the polynomial with polynomial from 1st to the 20th degree using the regression equation in the Python scikit-learn library, eventually storing the result in a csv format file [28].

Figure 3. Flowchart of the tool



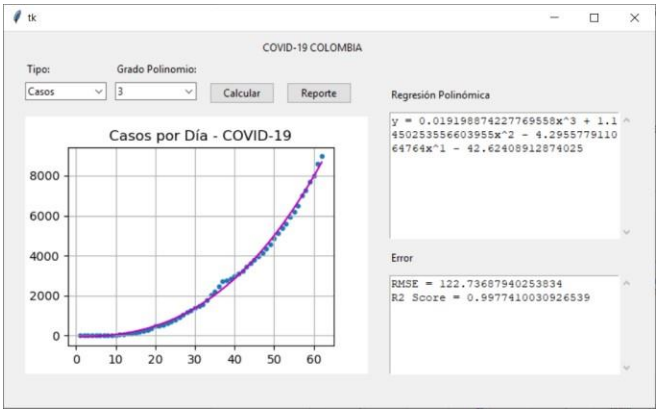
Source: Prepared by the authors

Based on the flowchart shown in Figure 3, Figure 4 denotes the graphical interface of the proposed tool. The tool has two selection lists (type and degree of polynomial), through which the user can choose the study type to be performed (cases, deaths, and recovered people) and the polynomial degree to be assessed through the tool. In Figure 4, we selected "cases" as the study type, and a polynomial degree of 5. After the user selected these parameters and pressed the "Calculate" button, the equation representing the polynomial regression for the study type and degree selected was obtained, which is shown in the upper right text area. Hence, for example, for the "cases" study type and a polynomial degree of 3, Equation (6) is obtained as follows:

$$y = 0.019198874227769558x^3 + 1.1450253556603955x^2 - 4.295577911064764x - 42.62408912874025 \quad . \quad (6)$$

Once the equation is obtained, the tool obtains the value of the quadratic RMSE and R^2 for the study type and the polynomial degree selected. For the "cases" study type and a polynomial degree of 3, the RMSE value is 122.7368 and the R^2 value is 0.9977.

Figure 4. The proposed tool



Source: Prepared by the authors

After performing the calculations of the regression equation and its assessment, the tool generates a graph in the left panel that includes the scattering of the loaded points and the trend line for the study type and the polynomial degree selected. The figure in the left panel is obtained using the functionalities provided by the Python matplotlib library. Finally, when the user presses the "Report" button, the tool generates a .csv report with the equations, RMSE, and R^2 of the polynomials of degrees 1–20 for the type of case selected. Figure 5 shows the general report for the "cases" study type.

https://doi.org/10.17081/invinno.8.2.4103

Figure 5. Generated .csv report

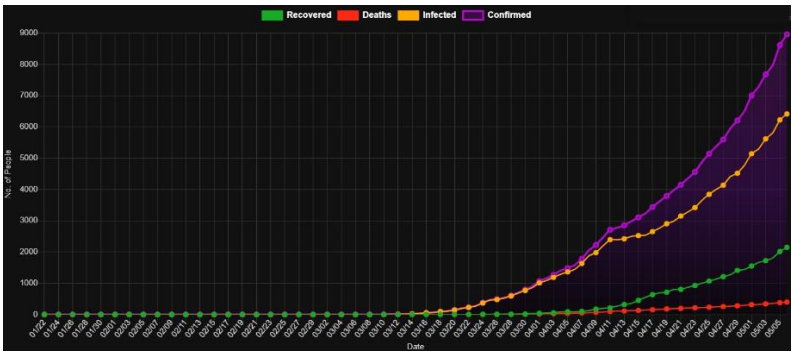
	A	B	C	D
1	grado	ecuac	rmse	r2
2	1	y = 136,054(	860,52015	0,8889579
3	2	y = 2,95931(	150,04804	0,9966238
4	3	y = 0,01919(	122,73688	0,997741
5	4	y = 0,00111(	94,590027	0,9986583
6	5	y = 6,33517(	64,759498	0,9993711
7	6	y = -4,1849(	64,374515	0,9993786
8	7	y = -1,15368	56,949869	0,9995136
9	8	y = -1,20245	56,744413	0,9995172
10	9	y = 2,00731(	53,914089	0,9995641
11	10	y = 5,08140(	52,841684	0,9995813
12	11	y = 1,75387(	52,178911	0,9995917
13	12	y = 4,53110(	55,29387	0,9995415
14	13	y = 1,23912(	60,685945	0,9994477
15	14	y = -4,83283	90,658621	0,9987675
16	15	y = -1,65354	117,02585	0,9979463
17	16	y = -4,98401	152,45098	0,9965148
18	17	y = -1,37307	194,83426	0,9943076
19	18	y = -3,53794	241,67722	0,9912414
20	19	y = -8,66035	290,77878	0,9873208
21	20	y = -2,03642	340,43818	0,9826203

Source: Prepared by the authors

Results

This section discusses the results obtained after performing the polynomial regression from the 1st to the 20th degree on the data of cases, deaths, and recovered people. Hence, Table 1 shows the polynomial degree, RMSE, and R^2 of each polynomial for the study type cases. These results were obtained by executing the report option from the proposed tool and using the functionalities provided by the scikit-learn, numpy, and matplotlib machine learning libraries. The data considered in the regression analysis for this case were from March 6, 2020 (the day of the first reported infection), to May 7, 2020, as shown in Figure 6.

Figure 6. COVID-19 curve in Colombia



Source: Taken from [29]

According to the results presented in Table 1, it can be concluded that, for the "cases" study type, out of the 20 polynomials assessed, the polynomial with the lowest RMSE and the highest R^2 is that of degree 11, followed by that of degree 10, which means that for the "cases" study type, these polynomials best represent the data curve.

Similarly, the regression that yields a higher RMSE is that of the polynomial with degree 1 (simple linear regression). Equation (7) shows the polynomial that best represents the data for the "cases" study type, and Figure 7 shows the curve generated by the tool for this study type.

Table 1. RMSE and R^2 for the "cases" study type

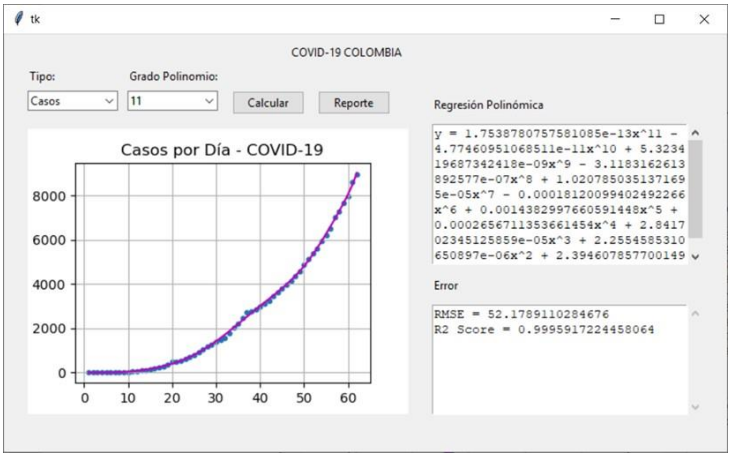
Order	RMSE	R^2
1	860.5201517	0.888957924
2	150.0480377	0.996623816
3	122.7368794	0.997741003
4	94.59002664	0.998658298
5	64.75949779	0.999371113
6	64.37451515	0.999378568
7	56.94986928	0.999513648
8	56.74441271	0.999517151
9	53.91408897	0.999564117
10	52.84168399	0.999581285
11	52.17891103	0.999591722
12	55.29386968	0.999541521
13	60.68594485	0.999447742
14	90.65862054	0.998767509
15	117.0258518	0.997946337
16	152.4509837	0.996514814
17	194.8342552	0.994307591
18	241.6772189	0.991241355
19	290.778783	0.987320827
20	340.4381775	0.982620311

Source: Prepared by the authors

$$Y = 1.7538 * 10^{-13}x^{11} - 4.7746 * 10^{-11}x^{10} + 5.3234 * 10^{-9}x^{9} - 3.1183 * 10^{-7}x^{8} + 1.0207 * 10^{-5}x^{7} - 0.0001x^{6} + 0.0014x^{5} + 0.00026x^{4} + 2.8417 * 10^{-5}x^{3} + 2.2554 * 10^{-6}x^{2} + 2.3946 * 10^{-5}x - 2.488 \quad . \quad (7)$$

https://doi.org/10.17081/invinno.8.2.4103

Figure 7. Curve of the "cases" study type



Source: Prepared by the authors

Table 2 indicates the polynomial degree, RMSE, and R^2 of each polynomial for the "deaths" study type. The data considered in the regression analysis for this case were from March 22, 2020 (the day of the first two reported deaths), to May 7, 2020.

According to the results presented in Table 2, it can be concluded that for the "deaths" study type, of the 20 polynomials assessed, the polynomial with the lowest RMSE and the highest R^2 is that of degree 10, followed by that of degree 9, which means that for the "deaths" study type, these polynomials best represent the data. Similarly, the regression that yields a higher RMSE is that of the polynomial with degree 1 (simple linear regression). Equation (8) shows the polynomial that best represents the data for deaths, and Figure 8 shows the curve generated by the tool for this study type.

Table 2. RMSE and R^2 of the "deaths" study type

Order	RMSE	R^2
1	25.91040964	0.953905931
2	7.765013976	0.995860185
3	6.657064706	0.99695728
4	5.001269793	0.998282657
5	3.217320227	0.999289302
6	2.851928235	0.999441564
7	2.585239027	0.999541121
8	2.532089425	0.999559795
9	2.540729209	0.999556786
10	2.529539736	0.999560681

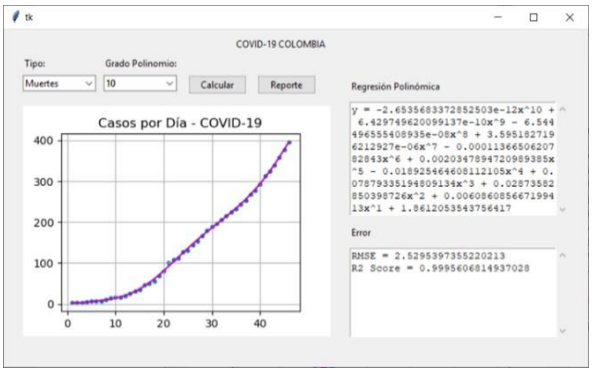
Order	RMSE	R^2
11	2.584017156	0.999541555
12	2.763525078	0.999475648
13	2.955807532	0.999400142
14	3.550107211	0.999134675
15	4.783814481	0.998428751
16	6.585864348	0.997022018
17	8.763384714	0.994727213
18	18.80550413	0.975719007
19	21.67532611	0.967742718
20	24.46572503	0.958902759

Source: Prepared by the authors

$$Y = -2.6535 * 10^{-12}x^{10} + 6.4297 * 10^{-10}x^{9} - 6.5444 * 10^{-8}x^{8} + 3.5951 * 10^{-6}x^{7} - 0.00011x^{6} + 0.002x^{5} - 0.0189x^{4} + 0.0787x^{3} + 0.0287x^{2} + 0.006x + 1.861 \quad (8)$$

Further, Table 3 denotes the polynomial degree, RMSE, and R^2 of each polynomial for the "recovered people" study type. The data considered in the regression analysis for this case were from March 17, 2020 (the day on which the first recovered person was reported), to April 20, 2020.

Figure 8. Curve of the "deaths" study type



Source: Prepared by the authors

Table 3. RMSE and R^2 for the "recovered people" study type

Order	RMSE	R^2
1	245.6592478	0.848932469
2	40.2717472	0.995940189
3	37.60720212	0.996459644
4	36.49088979	0.996666705

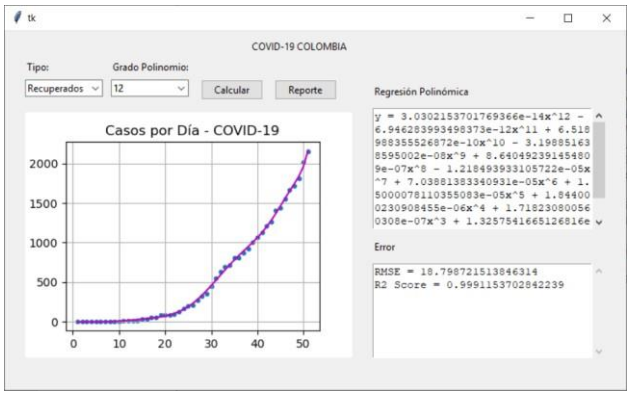
Order	RMSE	R^2
5	29.60697215	0.997805716
6	24.16200969	0.998538594
7	23.82103114	0.99857955
8	22.60964052	0.998720347
9	22.57039721	0.998724785
10	20.01924186	0.998996771
11	19.58884614	0.999039444
12	18.79872151	0.99911537
13	19.87213183	0.999011461
14	20.93223833	0.998903178
15	21.17405907	0.998877689
16	20.99236315	0.998896868
17	21.37937077	0.998855819
18	23.88041817	0.998572458
19	29.45398642	0.997828334
20	37.75312971	0.996432115

Source: Prepared by the authors

According to the results shown in Table 3, it can be concluded that for the "cases" study type, out of the 20 polynomials assessed, the polynomial with the lowest RMSE and the highest R^2 is that of degree 12, followed by that of degree 13, i.e., for the "cases" study type, these polynomials best represent the data. The regression that yields a higher RMSE is that of the polynomial with degree 1 (simple linear regression). Equation (9) shows the polynomial obtained for recovered people that best represents the data, and Figure 9 shows the curve generated by the tool for this study type.

$$Y = 3.0302 * 10^{-14}x^{12} - 6.9462 * 10^{-12}x^{11} + 6.5189 * 10^{-10}x^{10} - 3.198 * 10^{-8}x^{9} + 8.6404 * 10^{-7}x^{8} - 1.2184 * 10^{-5}x^{7} + 7.0388 * 10^{-5}x^{6} + 1.5 * 10^{-5}x^{5} + 1.844 * 10^{-6}x^{4} + 1.718 * 10^{-7}x^{3} + 1.325 * 10^{-8}x^{2} - 7.044 * 10^{-7}x - 1.098 \quad (9)$$

Figure 9. Curve of the "recovered people" study type.



Source: Prepared by the authors

The curves shown in Figures 7–9 accurately follow the data obtained for COVID-19 cases, deaths, and recovered people. In the long term, they present values that grow or decrease rapidly; hence, their greatest use is in obtaining an equation, which assesses the progress that the pandemic has had from an initial date to the current date. In this sense, if a long-term estimate should be obtained, simple linear regression or the polynomial of degree 1 may be more suitable. Hence, Table 4 shows estimates according to linear regression for the different studies: cases, deaths, and recovered people.

Table 4. Estimates for cases, deaths, and recovered people

Study	Equation	Prediction	Prediction
Cases	y = 136.05461962680366x − 1779.7043892120569	x = 50, y = 5023.03	x = 90, y = 10465.21
		x = 60, y = 6383.57	x = 100, y = 11825.76
		x = 70, y = 7744.12	x = 150, y = 18628.49
		x = 80, y = 9104.67	x = 200, y = 25431.22
Deaths	y = 8.878507554733268x − 68.84057971014482	x = 50, y = 375.08	x = 90, y = 730.23
		x = 60, y = 463.87	x = 100, y = 819.01
		x = 70, y = 552.65	x = 150, y = 1262.94
		x = 80, y = 641.44	x = 200, y = 1706.86
Recovered people	y = 39.5628959276018x − 476.6549019607842	x = 50, y = 1501.49	x = 90, y = 3084.01
		x = 60, y = 1897.12	x = 100, y = 3479.63
		x = 70, y = 2292.75	x = 150, y = 5457.78

https://doi.org/10.17081/invinno.8.2.4103

Study	Equation	Prediction	Prediction
		x = 80, y = 2688.38	x = 200, y = 7435.92

Source: Prepared by the authors

As for the COVID-19 cases, it is estimated that by day 200 of the pandemic, September 21, 2020, the number of infected people (cases) will reach 25,431. Regarding the number of deaths, it is estimated that by day 200 day of deaths, October 7, 2020, 1,706 deaths will occur. Finally, as for the number of recovered people, it is estimated that by day 200 after the first recovery, October 2, 2020, a total of 7,435 people will recover.

Conclusions

The linear and polynomial regression techniques are mathematical tools that can be used to predict and model the behavior of a dataset. Hence, this study aims to assess the behavior of the curve by means of a set of defined equations based on different governmental measures.

The RMSE and R^2 values obtained reveal that polynomial equations better model the behavior of COVID-19 data in Colombia. Regarding exponential approaches, however, given the sporadic behavior for future values, the equations obtained are particularly useful to assess the impact of government measures up to the date of data capture. In terms of prediction, linear regression may be more appropriate because the curve has a consistent behavior over time.

The machine learning libraries used in the development of the tool for calculating polynomial and linear regressions turned out to be suitable compared to commercial tools, as they extend the study to higher-degree polynomials, for directly obtaining RMSE and R^2. Hence, we could seize the advantages provided using the Python language and its set of libraries for machine learning (scikit-learn, numpy, and matplotlib).

In future works, we aim to perform comparative studies with the curves of other countries to identify the growth of Colombia with respect to other South American countries. We also intend to include other regression approaches in the tool.

References

1. Colombian Ministry of Health, ABC of the New Coronavirus (COVID-19). 2020 [Online]. Available: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RI DE/VS/PP/ET/abece-coronavirus.pdf.
2. UITP, COVID-19 Management - Guidelines for Public Transport Operators. 2020 [Online]. Available: https://www.uitp.org/sites/default/files/cck-focus-papers-files/Corona%20Virus_ESP.pdf.
3. H. L. Quach and N. A. Hoang, "COVID-19 in Vietnam: A lesson of pre-preparation," *Journal of Clinical Virology*, vol. 127, pp. 1-3, 2020.
4. M. Ruiz, "Health statistics and invisibility by sex and gender during the COVID-19 epidemic," *Sanitary Gazette*, 2020.
5. C. Hevia and A. Neumeyer, "A conceptual framework for analyzing the economic impact of COVID-19 and its repercussions on policies," United Nations Development Program, Buenos Aires, Argentina, 2020.
6. L. Ferreti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler – Dörner, M. Parker, D. Bonsall, and C. Fraser, "Quantifying SARS- CoV-2 transmission suggests epidemic control with digital contact tracing," *Science*, vol. 31, no. 6491, pp. 1-13, 2020.
7. A. Kucharski, T. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, and R. Eggo, "Early dynamics of transmission and control of COVID-19: A mathematical modelling study," *The Lancet Infectious Diseases*, vol. 20, no. 5, pp. 1-7, 2020.
8. Y. Vargas-Rodríguez, A. Obaya, and G. Vargas-Rodríguez, "Polynomial regression, an indispensable competence for data processing in chemical kinetics," *Contacts Magazine*, vol. 98, pp. 25-35, 2015.
9. H. Guerrero, *Excel Data Analysis - Modeling and Simulation*, Williamsburg: Springer, 2018.
10. E. Ostertagová, "Modelling using polynomial regression," *Procedia Engineering*, no. 48, pp. 500-506, 2012.
11. E. J. De la Hoz Domínguez, T. J. Fontalvo Herrera, and A. A. Mendoza, "Machine learning and SMEs: Opportunities for improving the decision-making process," *Research and Innovation in Engineering*, vol. 8, no. 1, pp. 21-26, 2020.

12. M. Janssen, H. van der Voort, and A. Wahyundi, "Factors influencing big data decision-making quality," *Journal of Business Research*, vol. 70, pp. 338-345, 2017.

13. L. Andrade and J. Cervantes, "Machine learning and classification models," *Polytechnic Magazine*, vol. 33, no. 1, 2014.

14. C. Soto-Valero, "Application of machine learning methods in the analysis and prediction of sports results," *New Trends in Physical Education, Sports and Recreation Magazine*, no. 34, pp. 377-382, 2018.

15. J. Cárdenas, G. Olivares, and R. Alfaro, "Automatic classification of texts using word networks," *Signs Magazine*, vol. 47, no. 86, pp. 346-364, 2014.

16. S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning from Theory to Algorithms*, New York: Cambridge University Press, 2014.

17. H. Ordoñez, C. Cobos, and V. Bucheli, "Machine learning model for the prediction of theft trends in Colombia," *RISTI Magazine*, vol. E29, pp. 494-506, 2020.

18. M. Kubat, *An Introduction to Machine Learning*, Miami - USA: Springer, 2017.

19. D. Barbery and K. Jury, "Impact of promotional spending in the traditional channel on the sales of the mass consumer company," *RISTI Magazine*, no. E26, pp. 15-28, 2020.

20. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, and M. Perrot, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

21. F. Dubosson, S. Bromuri, and M. Schumacher, "A python framework for exhaustive machine learning algorithms and features evaluations," in *2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), Crans-Montana - Switzerland*, 2016.

22. D. Paper, *Hands-on Scikit Learn for Machine Learning Applications*, Logan-USA: Springer, 2020.

23. S. Guido and A. Mueller, *Introduction to Machine Learning with Python: A Guide for Data Scientists*, Sebastopol: O'Reilly, 2016.

24. J. Hurwitz and D. Kirsch, *Machine Learning for Dummies*, Hoboken: John Wiley & Sons, 2018.

25. P. Vinuesa, “Simple and Polynomial Linear Regression: Theory and Practice,” 2016 [Online]. Available: https://www.ccg.unam.mx/~vinuesa/R4biosciences/docs/Tema 9_regresion.pdf.

26. D. Cardona, J. González, M. Rivera, and E. Cárdenas, “Application of linear regression in a poverty problem,” *Interaction Magazine*, vol. 12, pp. 73-84, 2013.

27. J. Cunningham, G. Valderrama, R. De Rodríguez, R. Sandoya, and M. Fernández, “Multiple regression model applied to the admission process of the University of Panama,” *Magazine of Mathematics: Theory and Applications*, vol. 14, no. 2, pp. 251-261, 2007.

28. J. Rawlings and S. D. D. Pantula, *Applied Regression Analysis*, New York: Springer, 1998.

29. COVID19Info, Covid19info.live. 2020 [Online]. Available: https://covid19info.live.