INVESTIGACIÓN e INNOVACIÓN en INGENIERÍAS
ISSN2344-8652

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

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

Gabriel Elías Chanchí Golondrino
Universidad de Cartagena, Colombia
Wilmar Yesid Campo Muñoz
Universidad del Quindío, Colombia
Luz Marina Sierra Martínez
Universidad del Cauca, Colombia

Open Access

Recibido:
21 de abril de 2020
Aceptado:
3 de julio de 2020
Publicado:
31 de julio de 2020

Correspondencia:
gchanchig@unicartagena.edu.co

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

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

Abstract

Objective: This research study proposes the use of the polynomial regression approach to model data behavior for characterizing behavior patterns for COVID-19 case, death, and recovery curves in Colombia. **Methodology:** Data were obtained from the reports provided by the Colombian Ministry of Health. First, the authors assessed the polynomial regression procedure through multiple linear regressions. Then, a tool was designed and implemented for applying this procedure on the collected dataset. Finally, the corresponding results were assessed. **Results:** For the three studies considered (cases, deaths, and recoveries), 20 polynomials were obtained together with the root mean square error (RMSE) and the determination level (R2) associated with each study. Further, 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. Therefore, they can be used to determine the impact of government measures over a given period of time. In addition, the machine learning libraries provided in Python language significantly supported the estimation of polynomial regressions.

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

Como citar (IEEE): G. Chanchí - Golondrino., W. Campo - Muñoz., y L. Sierra - Martínez, "Aplicación de la regresión polinomial para la caracterización de la curva del COVID-19, mediante técnicas de machine learning", Investigación e Innovación en Ingenierías, vol. 8, n°. 2, 2020. DOI: https://doi.org/10.17081/invinno.8.2.4103

Introducción

Los coronavirus (CoV) corresponden a virus que surgen de manera periódica en diferentes partes del mundo y que son causantes de la denominada Infección Respiratoria Aguda (IRA) con diferente nivel de gravedad. El COVID-19, en la actualidad se está transmitiendo a gran escala entre personas de diferentes regiones del mundo. La Organización Mundial de la Salud informó de la ocurrencia de diferentes casos de Infección Respiratoria Aguda Grave (IRAG) causada por el nuevo coronavirus (COVID-19) en Wuhan (China), desde la última semana de diciembre de 2019. Aunque se desconoce lo intensa que puede ser la transmisión, el contagio ocurre cuando una persona estornuda o tose expulsando gotículas que contienen el virus y entran en contacto con las personas cercanas o a través de las superficies en las que se hospedan [1, 2, 3, 4]. La pandemia del COVID-19 ha sido uno de los desafíos más complejos a los que se ha enfrentado la humanidad en la historia reciente, desconociéndose aun el costo total en vidas humanas y el impacto en diferentes sectores de la economía. El grado de contagio ha crecido a tal punto que ha generado una crisis en el sistema de salud de diferentes países, sumado al colapso económico que afectará de manera severa el bienestar de la población [3, 4, 5]. En el contexto particular de Colombia, el primer caso de infección de COVID-19 tuvo lugar el 6 de marzo de 2020, mientras que los dos primeros fallecimientos ocurrieron el 22 de marzo de 2020. El Ministerio de Salud ha estado emitiendo información a diario en los diferentes canales de comunicación sobre el número de casos, el número de fallecidos y el número de personas recuperadas por COVID-19. Estos reportes, actualizan la información de acuerdo a las pruebas que se van realizando, pero no incluyen estudios sobre la caracterización del comportamiento de los datos utilizando regresiones polinomiales.

La regresión polinomial puede ser considerada un caso particular de la regresión lineal múltiple, en la que se busca determinar el mejor polinomio que represente los datos de un conjunto de puntos [6, 7, 8, 9, 10]. En el contexto de Colombia, la obtención de un conjunto de polinomios que estimen el crecimiento de las curvas de casos, muertos y personas recuperadas, puede ser de gran interés para la toma de decisiones por parte de los entes gubernamentales. El aprendizaje automático apoya a los usuarios para realizar investigaciones, identificar patrones, detectar anomalías, y abogar por nuevos procedimientos [11, 12, 13, 14, 15]. En este mismo sentido, gracias a las herramientas actuales de aprendizaje automático (machine learning), es posible aplicar algunos algoritmos sobre un conjunto de datos, como es el caso del algoritmo de regresión lineal, con el fin de obtener una predicción sobre el comportamiento futuro de los datos [16, 17, 18, 19].

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

Aunque existen herramientas que permiten la obtención de tendencias asociadas a la regresión polinomial como excel, estas herramientas tienen como limitante que, la estimación de la tendencia polinomial automática solo aborda hasta polinomios de orden 6. Del mismo modo, no es posible obtener de manera directa el valor del error cuadrático medio. A partir de lo anterior, en el presente trabajo se utilizaron herramientas de machine learning provistas por el lenguaje Python, como es el caso de la librería scikit-learn [17, 18], con el fin de personalizar y ampliar el estudio de la regresión polinomial hasta polinomios de orden 20 sobre los datos obtenidos del COVID-19 en Colombia.

De acuerdo a lo anterior, en este artículo aporta el desarrollo de un estudio basado en el uso del algoritmo de machine learning de regresión polinómica, para determinar el comportamiento del crecimiento de la curva de casos, muertes y personas recuperadas del COVID-19 en el contexto de Colombia. Como apoyo para el desarrollo del presente estudio, en este artículo se desarrolló una herramienta en el lenguaje Python, la cual usa de las librerías de machine learning scikit-learn, numpy y matplotlib, que permiten el análisis de los datos y la obtención de la ecuación de regresión polinómica para polinomios de diferente orden [19, 20, 21]. Los datos analizados en el presente estudio fueron obtenidos a partir de los reportes realizados por el Ministerio de Salud de Colombia.

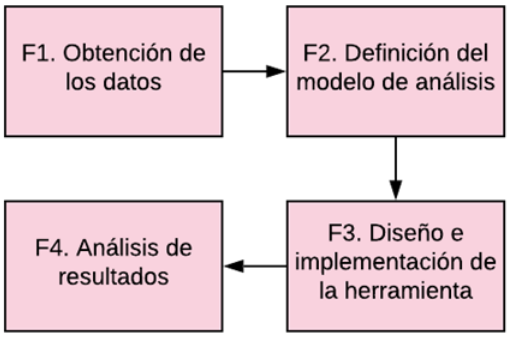
El estudio muestra, como las curvas de tendencia basadas en regresión polinómica, siguen de manera más fiel los datos que otros tipos de modelos matemáticos, sin embargo, dado que a largo plazo las ecuaciones polinómicas presentan una gran variación, la mayor utilidad de la regresión polinomial puede encontrarse en la evaluación de las medidas tomadas durante el periodo de cuarentena, para la toma de decisiones a futuro. Así mismo, este estudio busca servir de base para extrapolarlo a otros contextos de aplicación.

Metodología

Para el desarrollo de la presente investigación se tomaron en consideración las siguientes fases metodológicas: obtención de los datos, definición del modelo de análisis, diseño e implementación de la herramienta y análisis de resultados, como se muestra en la Figura 1.

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

Figura 1. Metodología propuesta



Fuente: Elaboración propia

En la primera fase, se obtuvieron los datos de los casos, muertes y personas recuperadas el COVID-19 en Colombia a partir del portal oficial y los canales de comunicación en redes sociales del Ministerio de Salud de Colombia. Estos datos fueron obtenidos desde el 6 de marzo de 2020, fecha en la que se confirmó el primer caso de contagio en Colombia. En la segunda fase, a partir de un análisis inicial de los datos y su comparación con otros modelos (exponencial y logarítmico), se definió hacer un estudio de regresión polinomial sobre los datos obtenidos en la fase 1, teniendo en cuenta el método de regresión lineal múltiple. En la fase 3, a partir de la definición del modelo, se diseñó e implementó una herramienta en el lenguaje Python para la conducción del estudio, la cual usa las librerías de machine learning scikit-learn, numpy y matplotlib. Finalmente, en la fase 4, a partir de la herramienta construida se aplicó el estudio de regresión polinómica para los datos asociados a los casos de contagio identificados, de muertes y personas recuperadas usando polinomios de orden 1 hasta orden 20.

Regresión polinomial

La regresión polinomial es un caso especial de la regresión lineal múltiple en la que se busca obtener la predicción de una variable de respuesta cuantitativa a partir de una variable predictora cuantitativa, donde la relación se modela como una función polinomial de orden o grado [22, 23, 24]. Mediante la regresión polinomial es posible obtener un polinomio como el presentado en la ecuación 1, en donde ε representa el error en la estimación o la diferencia entre el valor estimado y el valor observado.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots + \beta_p X^p + \varepsilon \quad (1)$$

El polinomio de la ecuación 1, puede ser expresado en el modelo lineal de la ecuación 2, el cual puede ser abordado como ya se mencionó, mediante una regresión lineal múltiple en la cual X1=X, X2=X2,….,Xp=Xp.

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

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p + \varepsilon \quad (2)$$

Para la obtención de la regresión lineal múltiple, es posible plantear el modelo matricial para n muestras de datos tal como se presenta en la ecuación 3.

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

De este modo, mediante la notación de matrices es posible expresar la ecuación 3, tal como se muestra en la ecuación 4.

$$Y = X\beta + \varepsilon \quad (4)$$

Dado que el objetivo de la ecuación matricial 4, es encontrar el vector de coeficientes β, es posible mediante propiedades matriciales (matriz traspuesta y matriz inversa) expresar la ecuación 4 en términos de la ecuación 5.

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

Así, el vector resultante β contiene los diferentes coeficientes $(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ del polinomio presentado en la ecuación 1. En este artículo, se hizo uso del anterior procedimiento matemático a través del uso de las funcionalidades provistas por la librería scikit-learn de Python, la cual aborda el problema como un caso especial de la regresión lineal.

Herramienta propuesta

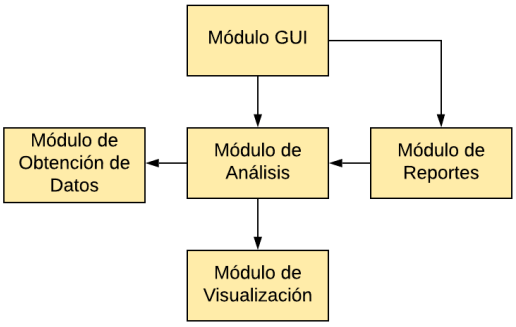
Este trabajo propone como aporte una herramienta para determinar el comportamiento del crecimiento de la curva de casos, muertes y personas recuperadas del COVID-19, usando para ello algoritmos de regresión polinómica mediante herramientas de machine learning. De este modo, en esta sección se presenta la estructura funcional y la implementación de la herramienta construida.

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

En cuanto al diseño de la herramienta propuesta, en la Figura 2 se presentan los diferentes módulos funcionales que la componen, dentro de los cuales se encuentran: módulo GUI, módulo de obtención de datos, módulo de análisis, módulo de visualización y módulo de reportes [25, 26, 27].

El módulo GUI, es el encargado de la gestión de los diferentes componentes gráficos de la interfaz (botones, cajas de texto, componentes gráficos), así como de la gestión de los diferentes eventos que involucran la interacción del usuario con dichos componentes. En el caso de la herramienta propuesta, este módulo fue implementado a partir de la librería Tkinter de Python, la cual permitió la construcción de la interfaz gráfica y la articulación de los diferentes componentes de interacción. Una vez son escogidas en la GUI de la herramienta las diferentes opciones a considerar para el análisis (tipo de estudio y grado del polinomio a considerar).

Figura 2. Módulos funcionales



Fuente: Elaboración propia

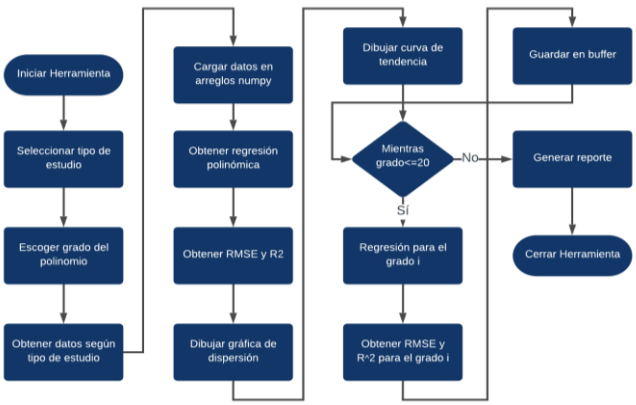
El módulo de análisis, obtiene y analiza los datos históricos de los casos de contagio, muertes y personas recuperadas del COVID-19, mediante el algoritmo de regresión polinómica, el cual toma en consideración el grado del polinomio y el tipo de estudio a realizar (casos, muertes, personas recuperadas) escogidos por el usuario. Los datos procesados por el módulo de análisis son obtenidos con apoyo del módulo de obtención de datos, el cual carga los datos desde archivos de texto plano y los transforma en arreglos propios de la librería numpy de Python. Para la obtención de la regresión polinómica, el módulo de análisis usa las funcionalidades provistas por la librería Python de machine learning scikit-learn. Una vez los datos son analizados, se evalúa el error cuadrático medio (RMSE) y el nivel de determinación (R2) de la curva de tendencia con respecto a los datos obtenidos (cercanía de la tendencia a los datos) y con el módulo de visualización, se genera una gráfica con los puntos originales y la curva de tendencia obtenida, proceso que es realizado mediante la librería matplotlib de Python. Finalmente, con el módulo de reportes, el usuario puede generar

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

un reporte comparativo con las diferentes tendencias por tipo de estudio (casos, muertes, personas recuperadas), teniendo en cuenta los diferentes grados de polinomio considerados en la herramienta y el error de la línea de tendencia con respecto a los datos asociado a cada grado.

En la Figura 3, se presenta un diagrama de flujo que ilustra el proceso seguido por la herramienta para realizar el análisis de los datos del COVID-19 y la obtención de las curvas de tendencia. En primera instancia, una vez iniciada la herramienta, el usuario escoge el tipo de estudio a realizar (casos, muertes y personas recuperadas) y el grado del polinomio a evaluar, es decir, el grado del polinomio que se usará para determinar la tendencia de los datos. A partir de lo anterior, la herramienta se encarga de obtener el conjunto de datos teniendo en cuenta el tipo de estudio y cargarlos en arreglos propios de la librería numpy de Python. En segunda instancia, la herramienta se encarga de aplicar la regresión polinómica sobre los datos dependiendo del tipo de grado del polinomio que haya escogido el usuario, obteniendo como resultado una ecuación que representa el comportamiento de los datos. Adicionalmente, la herramienta calcula la ecuación de tendencia el error cuadrático medio (RMSE) y el coeficiente de determinación (R2), que muestra que tan cercana es la tendencia a los datos analizados. Tanto la regresión como la evaluación de la misma es realizada a partir de las funcionalidades provistas por la librería de machine learning scikit-learn de Python. En tercera instancia, se genera una gráfica de dispersión con los datos cargados y la curva de tendencia que representa la ecuación obtenida para el grado del polinomio escogido, haciendo uso para ello de la librería matplotlib de Python. Finalmente, el usuario puede generar un reporte del tipo de estudio seleccionado, para lo cual la herramienta calcula mediante la librería scikit-learn de Python la ecuación de regresión, el error cuadrático medio y el nivel de determinación sobre los datos cargados para el polinomio con grado 1 hasta el polinomio con grado 20, almacenando el resultado en un archivo tipo csv [28].

Figura 3. Diagrama de flujo de la herramienta



Fuente: Elaboración propia

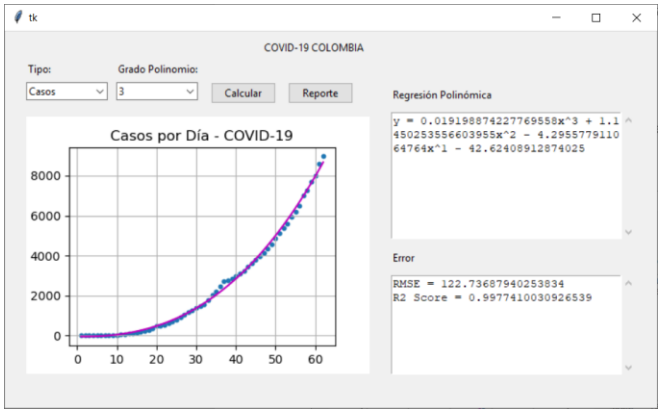
DOI: https://doi.org/10.17081/invinno.8.2.4103

A partir del diagrama de flujo presentado en la Figura 3, en la Figura 4 se muestra la interfaz gráfica de la herramienta propuesta. La herramienta cuenta con dos listas de selección (Tipo y Grado de Polinomio), a través de las cuales el usuario puede escoger el tipo de estudio que desea realizar (casos, muertes, personas recuperadas) y el grado del polinomio que desea evaluar a través de la herramienta. De este modo, en la Figura 4, se ha escogido como tipo de estudio: casos y como grado del polinomio: 5. Una vez seleccionado lo anterior por parte del usuario, al presionar el botón calcular se obtiene la ecuación que representa la regresión polinómica para el tipo de estudio y grado seleccionado, la cual se muestra en el área de texto de la parte superior derecha. Así a modo de ejemplo, para el tipo de estudio: casos y el grado de polinomio 3, se obtiene la ecuación 6:

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

Una vez obtenida la ecuación, la herramienta obtiene el valor del error cuadrático medio y el nivel de determinación para el tipo de estudio y el grado de polinomio seleccionado. Para el tipo de estudio casos y grado de polinomio 3, el valor del RMSE es de 122.7368 y el valor de R2 es de 0.9977.

Figura 4. Herramienta propuesta



Fuente: Elaboración propia

Después de realizados los cálculos de la ecuación de regresión y la evaluación de la misma, la herramienta genera en el panel de la izquierda una gráfica que incluye la dispersión de los puntos cargados y la línea de tendencia para el tipo de estudio y el grado del polinomio seleccionados. La Figura cargada en el panel de la izquierda es obtenida mediante las funcionalidades provistas por la librería matplotlib de Python. Finalmente, cuando el usuario presiona el botón reporte, la herramienta genera un reporte csv con las ecuaciones, el error cuadrático medio y el nivel de determinación de los polinomios del grado 1 al 20 para el tipo de caso

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

seleccionado. En la Figura 5, se muestra el reporte general para el tipo de estudio: casos.

Figura 5. Reporte CSV generado

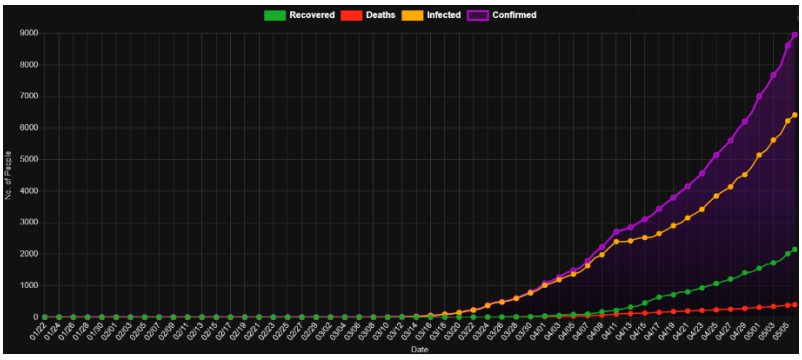
	A	B	C	D
1	grado	ecuac	rmse	r2
2	1	y = 136,054(	860,52015	0,8889579
3	2	y = 2,959311	150,04804	0,9966238
4	3	y = 0,019191	122,73688	0,997741
5	4	y = 0,001115	94,590027	0,9986583
6	5	y = 6,335172	64,759498	0,9993711
7	6	y = -4,18496	64,374515	0,9993786
8	7	y = -1,15368	56,949869	0,9995136
9	8	y = -1,20245	56,744413	0,9995172
10	9	y = 2,007318	53,914089	0,9995641
11	10	y = 5,081406	52,841684	0,9995813
12	11	y = 1,753878	52,178911	0,9995917
13	12	y = 4,531101	55,29387	0,9995415
14	13	y = 1,239127	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

Fuente: Elaboración propia

Resultados

En esta sección se presentan los resultados obtenidos luego de realizar la regresión polinomial desde el orden 1 hasta el orden 20 sobre los datos de los casos, muertes y personas recuperadas. De este modo en la Tabla 1, se muestra el orden del polinomio, el error cuadrático medio (RMSE) y el nivel de determinación (R2) de cada polinomio para el tipo de estudio casos. Estos resultados fueron obtenidos al ejecutar la opción reporte desde la herramienta propuesta y haciendo uso de las funcionalidades provistas por las librerías de machine learning scikit-learn, numpy y matplotlib. Los datos considerados en el análisis de regresión para este caso fueron tomados desde el 6 de marzo de 2020 (día del primer contagio reportado) hasta el 7 de mayo de 2020, tal como se muestra en la Figura 6.

Figura 6. Curva del COVID-19 para Colombia



Fuente: Tomado de [29]

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

De acuerdo a los resultados presentados en la Tabla 1, se puede concluir que para el tipo de estudio casos, de los 20 polinomios evaluados, el polinomio con el menor error cuadrático medio y mayor nivel de determinación es el de grado 11, seguido por el de grado 10, lo que quiere decir que para el tipo de estudio casos, estos polinomios son los que mejor representan la curva de los datos.

Del mismo modo, la regresión que arroja un error cuadrático más alto es la del polinomio con grado 1 (regresión lineal simple). En la ecuación 7, se presenta el polinomio que mejor representa los datos para el tipo de estudio casos y en la Figura 7 se presenta la curva generada por la herramienta para este tipo de estudio.

Tabla 1. RMSE y R^2 para el tipo de estudio casos

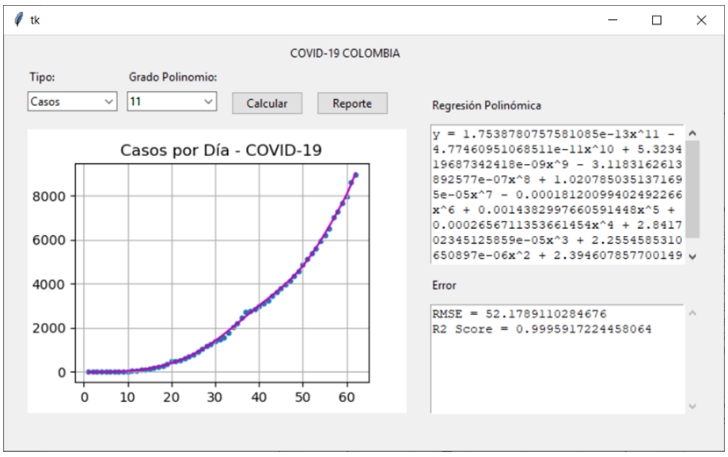
Orden	RMSE	R^2
1	860.5201517	0.888957924
2	150.0480377	0.996623816
3	122.7368794	0.997741003
4	94.59002664	0.998658298
5	64.75949779	0.999371113
6	64.37451515	0.999378568
7	56.94986928	0.999513648
8	56.74441271	0.999517151
9	53.91408897	0.999564117
10	52.84168399	0.999581285
11	52.17891103	0.999591722
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

Fuente: Elaboración propia

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

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

Figura 7. Curva para el tipo de estudio casos



Fuente: Elaboración propia

En la Tabla 2, se muestra el orden del polinomio, el error cuadrático medio (RMSE) y el nivel de determinación (R2) de cada polinomio para el tipo de estudio muertes. Los datos considerados en el análisis de regresión para este caso fueron tomados desde el 22 de marzo de 2020 (día de las primeras dos muertes reportadas) hasta el 7 de mayo de 2020.

De acuerdo a los resultados presentados en la Tabla 2, se puede concluir que para el tipo de estudio muertes, de los 20 polinomios evaluados, el polinomio con el menor error cuadrático medio y mayor nivel de determinación es el de grado 10, seguido por el de grado 9, lo que quiere decir que para el tipo de estudio muertes, estos polinomios son los que mejor representan los datos. Del mismo modo la regresión que arroja un error cuadrático más alto es la del polinomio con grado 1 (regresión lineal simple). En la ecuación 8, se presenta el polinomio que mejor representa los datos de muertes y en la Figura 8 se presenta la curva generada por la herramienta para este tipo de estudio.

Tabla 2. RMSE y R2 para el tipo de estudio muertes

Orden	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

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

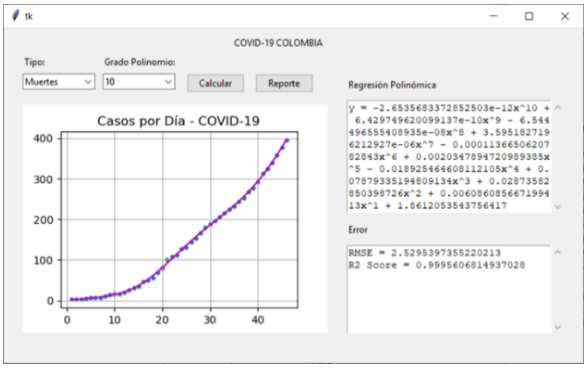
Orden	RMSE	R^2
11	2.584017156	0.999541555
12	2.763525078	0.999475648
13	2.955807532	0.999400142
14	3.550107211	0.999134675
15	4.783814481	0.998428751
16	6.585864348	0.997022018
17	8.763384714	0.994727213
18	18.80550413	0.975719007
19	21.67532611	0.967742718
20	24.46572503	0.958902759

Fuente: Elaboración propia

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

En la Tabla 3, por su parte, se muestra el orden del polinomio, el error cuadrático medio (RMSE) y el nivel de determinación (R2) de cada polinomio para el tipo de estudio personas recuperadas. Los datos considerados en el análisis de regresión para este caso fueron tomados desde el 17 de marzo de 2020 (día en el que se reportó la primera persona recuperada) hasta el 20 de abril de 2020.

Figura 8. Curva para el tipo de estudio muertes



Fuente: Elaboración propia

Tabla 3. RMSE y R^2 para el tipo de estudio personas recuperadas

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

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

Orden	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

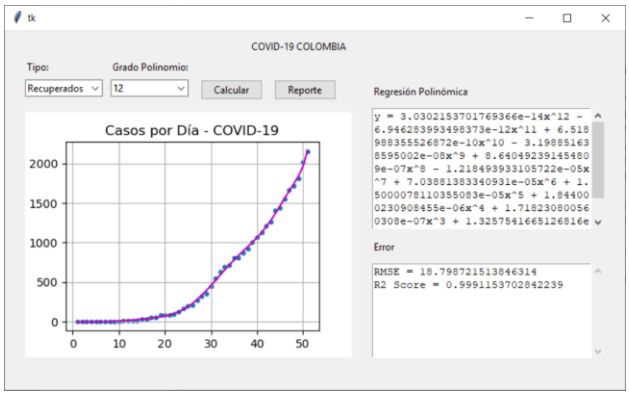
Fuente: Elaboración propia

De acuerdo a los resultados presentados en la Tabla 3, se puede concluir que para el tipo de estudio casos, de los 20 polinomios evaluados, el polinomio con el menor error cuadrático medio y mayor nivel de determinación es el de grado 12, seguido por el de grado 13, es decir, para el tipo de estudio de casos, estos polinomios son los que mejor representan los datos. La regresión que arroja un error cuadrático más alto es la del polinomio con grado 1 (regresión lineal simple). En la ecuación 9, se presenta el polinomio obtenido de personas recuperadas que mejor representa los datos y en la Figura 9 se presenta la curva generada por la herramienta para este tipo de estudio.

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

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

Figura 9. Curva para el tipo de estudio Recuperados



Fuente: Elaboración propia

Las curvas visualizadas en las Figuras 7, 8 y 9, siguen de manera fiel los datos obtenidos con respecto a los casos, muertes y personas recuperadas, a largo plazo presentan valores que crecen o decrecen rápidamente, por lo que su mayor utilidad es en la obtención de una ecuación, que permite analizar el progreso que ha tenido la pandemia desde una fecha inicial hasta la fecha actual. En este sentido, si se requiere obtener una estimación a largo plazo, la regresión lineal simple o el polinomio de grado 1 puede ser más adecuado. Así, en la Tabla 6, se presentan estimaciones de acuerdo a la regresión lineal para los diferentes estudios: casos, muertes y personas recuperadas.

Tabla 6. Estimaciones para los casos, muertes y personas recuperadas

Estudio	Ecuación	Predicción	Predicción
Casos	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
Muertes	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
Recuperados	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

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

Estudio	Ecuación	Predicción	Predicción
		$x=80$, $y = 2688.38$	$x=200$, $y = 7435.92$

Fuente: Elaboración propia

En cuanto a los casos de COVID-19, se estima que para el día 200 de iniciado el contagio se haya alcanzado el número de 25431 de infectados (casos), para el 21 de septiembre de 2020. En lo referente al número de muertes, se estima que para el día 200 de iniciados los fallecimientos se haya alcanzado un total de 1706 muertes, es decir, para el 7 de octubre de 2020. Finalmente, en cuanto al número de recuperados, se estima que para el día 200 después de iniciadas las recuperaciones, se haya alcanzado un total de 7435 personas recuperadas, para el 2 de octubre de 2020.

Conclusiones

Las técnicas de regresión lineal y polinomial son herramientas matemáticas, que pueden ser utilizadas para predecir y modelar el comportamiento de un conjunto de datos, por lo que este estudio pretende ser de ayuda para evaluar el comportamiento de la curva mediante un conjunto de ecuaciones definidas, ante las diferentes medidas gubernamentales.

Los valores obtenidos de RSME y R2 permiten concluir que las ecuaciones polinomiales modelan de mejor forma el comportamiento de los datos del COVID-19 en Colombia, con respecto a los enfoques exponenciales, sin embargo, dado el comportamiento esporádico para valores futuros, las ecuaciones obtenidas son de mucha utilidad de cara a evaluar el impacto de las medidas gubernamentales hasta la fecha de captura de los datos. En lo referente a la predicción, la regresión lineal puede ser más adecuada dado que la curva tiene un comportamiento consistente en el tiempo.

Las librerías de machine learning empleadas en el desarrollo de la herramienta para el cálculo de las regresiones polinómicas y lineales, resultaron ser adecuadas con respecto a las herramientas comerciales, dado que permiten extender el estudio hasta polinomios de mayor grado, así como obtener de manera directa el RMSE y el R2. En este sentido se aprovecharon las ventajas provistas por el lenguaje Python y su conjunto de librerías para machine learning (scikit-learn, numpy, matplotlib).

Como trabajo futuro derivado de la presente investigación, se pretende realizar estudios comparativos con las curvas de otros países con el fin de identificar el crecimiento de Colombia con respecto a otros países de Suramérica. Así mismo, se pretende incluir en la herramienta otros enfoques de regresión.

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

Referencias bibliográficas

1. Ministerio de Salud de Colombia, «Abecé Nuevo Coronavirus (COVID-19),» 2020. [En línea]. Available: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/PP/ET/abece-coronavirus.pdf.

2. UITP, "Gestión de COVID-19 - Directrices para operadores de transporte público", 2020. [En línea]. 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, "Las estadísticas sanitarias y la invisibilidad por sexo y de género durante la epidemia de COVID-19", *Gaceta Sanitaria*, 2020.

5. C. Hevia y A. Neumeyer, "Un marco conceptual para analizar el impacto económico del COVID-19 y sus repercusiones en las políticas", Programa de las Naciones Unidas para el Desarrollo, 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, pp. 1-13, 2020.

7. A. Kucharski, T. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk y R. Eggo, "Early dynamics of transmission and control of COVID-19: a mathematical modelling study", *The Lancet Infectious Diseases*, pp. 1-7, 2020.

8. Y. Vargas-Rodríguez, A. Obaya y G. Vargas-Rodríguez, "Regresión polinómica, una competencia indispensable para el tratamiento de datos en cinética química" *Revista Contactos*, 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*, n°. 48, pp. 500-506, 2012.

11. E. J. De la Hoz Domínguez, T. J. Fontalvo Herrera y A. A. Mendoza Mendoza, "Aprendizaje automático y PYMES: Oportunidades para el mejoramiento del proceso de toma de decisiones" *Investigación e Innovación en Ingenierías*, vol. 8, n°. 1, pp. 21-26, 2020.

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

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

13. L. Andrade y J. Cervantes, "Aprendizaje automático y modelos de clasificación", *Revista Politécnica*, vol. 33, n°. 1, 2014.

14. C. Soto-Valero, "Aplicación de métodos de aprendizaje automático en el análisis y la predicción de resultados deportivos, *Revista Nuevas Tendencias en Educación Física, Deportes y Recreación*, n°. 34, pp. 377-382, 2018.

15. J. Cárdenas, G. Olivares y R. Alfaro, "Clasificación automática de textos usando redes de palabras", *Revista Signos*, vol. 47, n°. 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 y V. Bucheli, "Modelo de machine learning para la predicción de las tendencias de hurto en Colombia", *Revista RISTI*, vol. E29, pp. 494-506, 2020.

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

19. D. Barbery y K. Jurado, "Impacto del gasto promocional en el canal tradicional sobre las ventas de la empresa de consumo masivo", *Revista RISTI*, n°. 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 y M. Schumacher, "A Python Framework for Exhaustive Machine Learning Algorithms and Features Evaluations", *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 y D. Kirsch, *Machine Learning for Dummies*, Hoboken: John Wiley & Sons, 2018.

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

25. P. Vinuesa, "Regresión Lineal Simple y Polinomial: teoría y práctica", 2016. [En línea]. Available: https://www.ccg.unam.mx/~vinuesa/R4biosciences/docs/Tema9_regresion.pdf.

26. D. Cardona, J. González, M. Rivera y E. Cárdenas, "Aplicación de la regresión lineal en un problema de pobreza", *Revista Interacción*, vol. 12, pp. 73-84, 2013.

27. J. Cunningham, G. Valderrama, R. De Rodríguez, R. Sandoya y M. Fernández, "Modelo de Regresión Múltiple Aplicado al Proceso de Admisión de la Universidad de Panamá", *Revista de Matemática: Teoría y Aplicaciones*, vol. 14, n°. 2, pp. 251-261, 2007.

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

29. Covid19info.live, "covid19info.live", Enero 2020. [En línea]. Available: https://covid19info.live. [Último acceso: 5 Mayo 2020].

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