Modelos de evolución del Covid-19 en Panamá

Agapito Ledezma Espino* y José Javier Laguardia†
*Departamento de Informática
Universidad Carlos III de Madrid Madrid, España.
Email: agapito.ledezma@uc3m.es
†Universidad Tecnológica de Panamá, Panamá, Panamá
Email: jose.laguardia@utp.ac.pa

I. Introducción

A finales del año 2019 se reportaron numerosos casos de una nueva neumonía en Wuhan, capital de la provincia de Hubei, China. No fue hasta el 31 de diciembre cuando la Organización Mundial de la Salud (OMS) recibe la primera alerta de que se trataba de un nuevo tipo de Coronavirus. Posteriormente, este tipo de virus fue nombrado como nuevo Coronavirus 2019 (actualmente SARS-Cov-2) y la enfermedad infecciosa causada por el virus como COVID-19. Este virus ya ha causado una de las mayores crisis sanitarias del último siglo, con consecuencias no solo sanitarias, sino también económicas y sociales a las cuales Panamá no escapa.

El Ministerio de Salud de Panamá (MINSA), en línea con las recomendaciones de la Organización Mundial de la Salud, está implementando las medidas de prevención, detección temprana y control que permitan brindar una respuesta sanitaria integral necesaria para la atención y protección de la población susceptible de ser afectada por el virus Covid 19.

A pesar de las acciones que se están llevando a cabo, dada la magnitud de la pandemia y la gravedad de la misma, es necesario abordar el problema desde un punto de vista holístico. Por esta razón, a nivel mundial se están llevando a cabo diversas iniciativas orientadas a combatir la pandemia a partir de los datos.

En este informe hacemos una predicción a medio plazo del comportamiento de la enfermedad tomando en consideración las medidas tomadas por el gobierno de Panamá usando un modelo epimediológico, así como un pronostico a corto plazo usando un enfoque diferente basado en Redes Neuronas Artificiales, Arboles y Reglas de Regresión.

II. Modelo Epidemiológico SIR

El el año 1927, Kermack y McKendrick [1] formularon un modelo matemático bastante general y complejo denominado SIR donde se establecen las siguientes premisas:

- La enfermedad que iban a estudiar debía ser viral o bacteriana y ser transmitida por contacto directo de persona a persona.
- Al inicio de la epidemia solamente una fracción de la población era contagiosa.
- A excepción de las pocas personas inicialmente enfermas, todas las demás eran susceptibles de enfermarse.
- El individuo sufre el curso completo de la enfermedad para al final recuperarse adquiriendo inmunidad, o morir.
- Por último, la población total de personas se mantiene constante.

Siendo N la población total, la cual se mantiene constante a lo largo del problema y S(t), I(t) y R(t) los individuos susceptibles de ser enfermados, infectados y removidos de los infectados, es decir curados y fallecidos, el modelo se puede describir mediante el diagrama de la Figura 1.

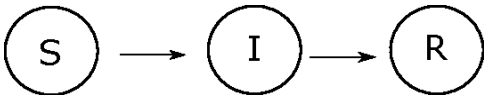

Fig. 1. Modelo SIR

De esta forma el modelo cumple:

$$\frac{dS}{dt} = -f(S,I),\ \frac{dI}{dt} = f(S,I) - g(I),\ \frac{dR}{dt} = g(I).$$

donde $f(S,I)$ es la incidencia de la enfermedad, es decir la tasa de infección. Se puede considerar

$$f(S,I) = \alpha(I)S = \beta IS$$

con $\alpha(I)$ la probabilidad de infección de una persona susceptible. De esta forma se considera $\alpha(I) = \beta I$ donde el parámetro β es la tasa de infección por contactos. La función $g(I)$ representa la probabilidad de que una persona infectada deje de estarlo. En este caso $g(I) = \gamma I$ de forma que γ representa la probabilidad de que un individuo se recupere en el siguiente intervalo de tiempo.

En general, se puede afirmar que el promedio del tiempo de permanencia de un individuo en la clase I es de $\frac{1}{\gamma}$ y que la tasa de reproducción es $R = \frac{\beta N}{\gamma}$ con la cual una persona infectada contagia a los demás individuos.

De esta forma el modelo quedaría.

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= \beta SI - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned}$$

III. Solución numérica

Para resolver el sistema de ecuaciones diferenciales ordinarias usamos en un primer momento Euler que podría verse como un sistema de diferencias finitas explicito hacia delante de incremento en el tiempo de un día:

$$\begin{aligned}\frac{S(t+1)-S(t)}{\triangle t} &= -\beta S(t)I(t)\\ \frac{I(t+1)-I(t)}{\triangle t} &= \beta S(t)I(t)-\gamma I(t)\\ \frac{R(t+1)-R(t)}{\triangle t} &= \gamma I(t)\end{aligned}$$

Para valores de $\beta = 0{,}5$ y $\gamma = 0{,}25$ y datos iniciales de $S(0) = 1$, $I(0) = 1{,}27 * 10^{-6}$ y $R(0) = 0$ se obtiene la solución mostrada en la figura 2

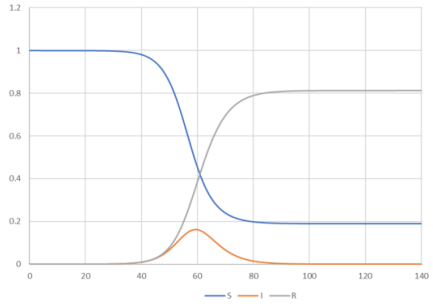

Fig. 2. Modelo epidemiológico SIR con $R_0 = 2$

Estos modelos tienen la ventaja de ser muy sencillos de implementar pero sólo ofrecen orden 1 en el tiempo. Por ese motivo se decidió utilizar una fórmula explícita de Runge-Kutta (4,5). En la Figura 3 se puede observar que la diferencia es significativa entre ambos, en el modelo resuelto mediante un Runge-Kutta con una aproximación de orden 4 se llega al pico de infectados antes pero con un número menor de casos.

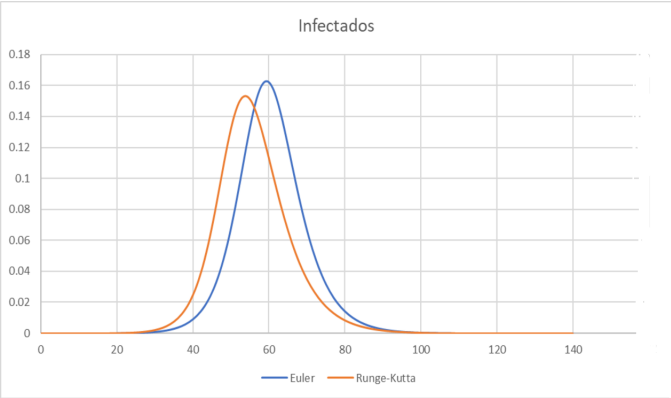

Fig. 3. Comparación entre el método de Euler y un Ringe-Kutta (4,5)

Sin embargo, un modelo así planteado no permite reflejar las medidas introducidos por las autoridades para combatir el Coronavirus. Por ese motivo, se introduce un modelo dependiente del tiempo como el planteado por Chen [2].

$$\begin{aligned}\frac{dS}{dt} &= -\beta(t)SI\\ \frac{dI}{dt} &= \beta(t)SI-\gamma I\\ \frac{dR}{dt} &= \gamma(t)I\end{aligned}$$

En este caso los valores de $\beta(t)$ tienen la forma de una exponencial $b_1 e^{-b_2 t}$ y los de $\gamma = \frac{(1-g_1^{g_2 t})}{dias}$. Donde días son los días promedio que un contagiado tarda en recuperarse y $g_2 < 0$

IV. Resultados del modelo para Hubei

A continuación, se presentan los datos que releja el modelo para Hubei, China considerando la proyección del 11 de febrero, antes del cambio de la forma de contar los casos positivos, y cuatro días antes, el 7 de febrero.

Primero, se puede ver el número de infectados totales en la figura 4. Se puede observar como la curva de infectados va bajando según se incrementa los datos disponibles. De hecho aunque la curva del 11 de febrero parece inferior a la real, esta aun se mantiene ligeramente superior si no se hubiera realizado el incremento de casos debido al cambio de definición de infectado.

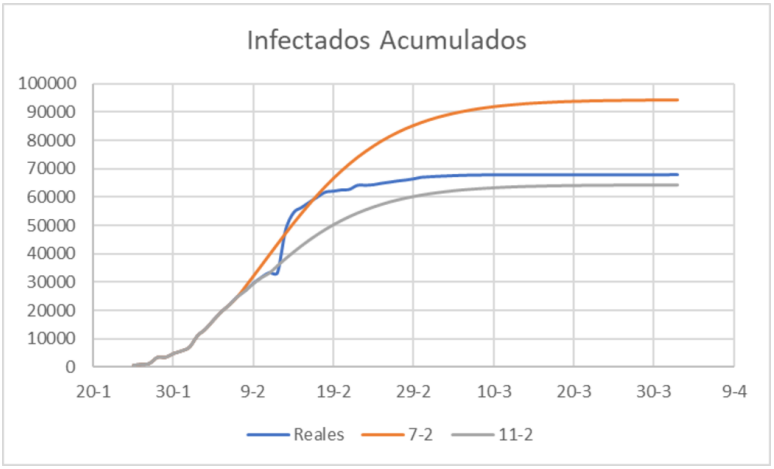

Fig. 4. Número de infectados acumulados en Hubei, China

Por tanto, en la Figura 5 se centra en la evolución del momento del pico. De forma que el 11 de febrero este se esperaba para el 20 de febrero cuando este fue el 18. Mientras que cuatro días antes se proyectaba un pico para el 22 de febrero.

Si bien es cierto que hay un cambio significativo en los casos finales que se muestran en la figura 4 esto se debe en gran medida al cambio de las medidas tomadas por China, de forma que el modelo parece que es capaz de ajustarse a estas tal como se buscaba.

V. Proyección para Panamá

Para el caso de Panamá calculamos una gráfica de βN y γ a partir de los datos dados por el MINSA. De esta forma se

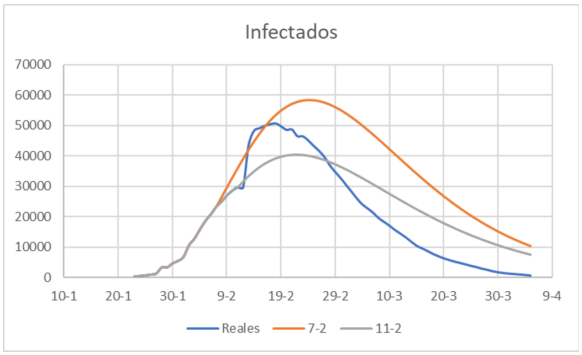

Fig. 5. Número de infectados en Hubei, China

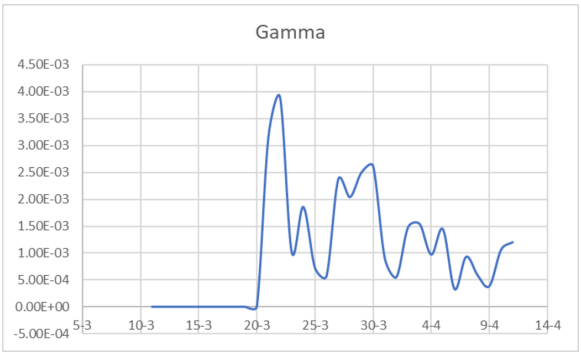

Fig. 7. Valores de para los primeros 32 días

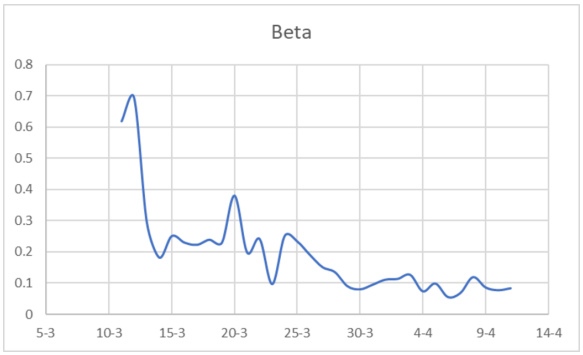

Fig. 6. Valores de para los primeros 32 días

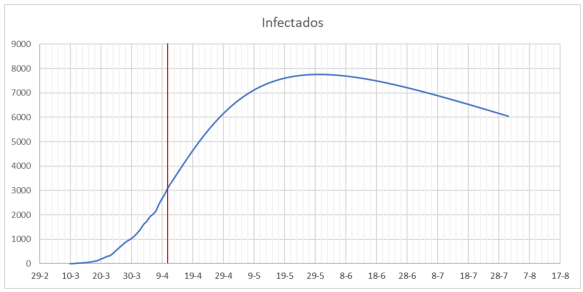

Fig. 8. Pronóstico de infectados

ajusta $\beta(t)$ mediante gradiente conjugado para ajustarse a la figura 6. Y el valor de $\gamma(t)$ a partir de los valores de la figura 7.

Aquí se puede ver un claro problema respecto a los valores de γ que son muy bajos. Esto se debe a la falta de registro por parte del MINSA de los recuperados. El Ministerio sólo cuenta aquellas personas que han dado negativo tras realizar una segunda prueba, Y únicamente se esta haciendo segundas pruebas a los hospitalizados que parecen recuperados. Toda la población que ha tenido síntomas leves y que a pasado la enfermedad en cuarentena domiciliaria se desconoce si se han recuperado. Ya que debido al limitado número de pruebas disponibles, esta información parece no prioritaria frente a identificar nuevos casos.

El comportamiento de las personas infectadas que se proyecta con los datos a 11 de abril se muestran en la Figura 8. Alcanzándose el pico el 29 de mayo. Igualmente se muestra el número de infectados totales en la figura 9, aquí se presenta una evolución de la estimación. Observamos como disminuye desde el 12 de abril hasta el 14, allí comienza de nuevo a aumentar. Sin embargo el aplanamiento se mantiene siempre en el mes de junio y nos da una estimación del rango en que nos movemos de infectados finales.

Finalmente en la siguiente tabla se muestran las proyecciones de personas infectadas de los últimos días versus el valor real. Aunque este modelo no pretende usarse para proyecciones a corto plazo.

Día	Proyectado	Real
10 abril	2936	2974
11 abril	3164	3234
12 abril	3433	3400
13 abril	3614	3472
14 abril	3663	3574
15 abril	3747	3754

Como se ve el modelo es muy sensitivo a la calidad de los datos, el 13 y 14 de abril las pruebas realizadas para detectar infectados se redujo a más de la mitad. Y aunque el modelo se adapta a todos estos cambios demora unos días en ajustarse.

El modelo es capaz de estimar el número de infectados detectados si se sigue la misma tendencia en pruebas realizadas con las medidas actuales de cuarentena. Cabe señalar que la curva mostrada en la Figura 8 muestra una pendiente muy baja una vez alcanzado el pico debido al bajo número de casos recuperados, siendo esta mucho más pronunciada en la realidad y que se espera poder obtener según aumenten los datos disponibles.

VI. Aproximación desde la Inteligencia Artificial y el Aprendizaje Automático

Como ya se ha comentado, la situación que ha provocado la pandemia causada por el Covid-19 debe abordarse desde distintos puntos de vista. La Inteligencia Artificial (IA) y, en particular, el Aprendizaje Automático (AA) han demostrado ser herramientas eficaces en generación de modelos guiados por los datos en diversos dominios. En general, el aprendizaje automático trata de la construcción de programas que, utilizando la experiencia sean capaces de mejorar automáticamente su rendimiento. Según Mitchell [3] *"un programa de ordenador*

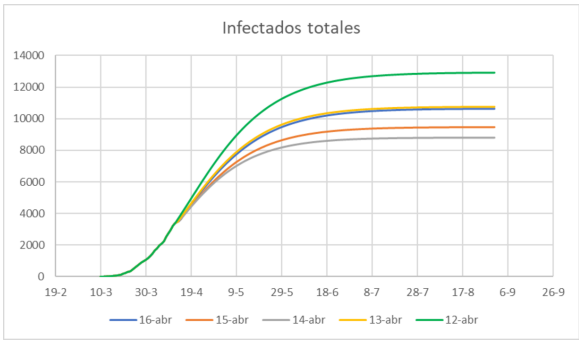

Fig. 9. Pronóstico de infectados acumulados

se dice que aprende de la experiencia E con respecto a una cierta clase de tarea T y medida de funcionamiento P, si su funcionamiento en la tarea T según lo medido por P, mejora con la experiencia E."

Los sistemas basado en técnicas de Aprendizaje Automático, al igual que todos los sistemas guiados por los datos (*data-drive*) son sistemas $GIGO$ (*Garbage in - Garbage out*) por lo que la *calidad de los datos* es fundamental para obtener buenos sistemas.

En las últimas semanas se han lanzado varias iniciativas a nivel mundial con el propósito de abordar los problemas causados por la pandemia provocada por el Covid-19 desde la perspectiva del Aprendizaje Automático y la Ciencia de los Datos [4]. Son diversas la formas en las que el Aprendizaje Automático puede ayudar a luchar contra la pandemia. El AA se podría utilizar, por ejemplo, para [5]:

- Identificar quién está en mayor riesgo
- Diagnosticar a los pacientes
- Desarrollar los medicamentos más rápido
- Predecir la propagación de la enfermedad
- Comprender mejor los virus
- Mapear de donde vienen los virus
- Predecir la próxima pandemia

En este trabajo preliminar se aborda el problema de la predicción de la propagación mediante modelos de regresión basado en Aprendizaje Automático. Para llevar a cabo la generación de los modelos se ha seguido la metodología de minería de datos CRISP-DM [6].

VI-A. Descripción de los datos

Para crear los modelos de evolución de la pandemia se han utilizado los datos de *The Humanitarian Data Exchange*, en particular, el *Novel Coronavirus (COVID-19) Cases Data* [1] que recopila datos epidemiológicos del COVID-19 desde el 22 de enero de 2020. Los datos son recopilados por el *Center for Systems Science and Engineering* de la Universidad Johns Hopkins (JHU CCSE) a partir de varias fuentes.

El conjunto de datos utilizados tiene un total de 264 registros que corresponden a Provincias/Estados de distintos países/regiones. Como ya se ha comentado, el conjunto de datos recopila el número de infectados totales desde el día 22 de enero hasta la fecha. A día de hoy (15 de abril) hay un total de 88 atributos (columnas) por cada registro. Los primeros cuatro atributos corresponden al nombre de la provincia/estado, la región/país y la longitud y latitud del mismo. El resto de atributos reflejan el número de contagiados acumulados por día.

[1]https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases

VI-B. Creación de modelos

Para el pre-procesado de datos y la creación de los modelos se ha utilizado la herramienta de software libre para el análisis de conocimiento Weka [7]. Siguiendo la metodología CRISP-DM, se han llevado a cabo varias iteraciones con el propósito de encontrar el modelo que mejor se ajuste a los datos. Para ello, se han utilizado como parámetros de evaluación el coeficiente de correlación, el error absoluto relativo y la raíz del error cuadrático relativo. Todos los modelos son evaluados a través de un proceso de validación cruzada de 10 iteraciones. Para generar los modelos se han utilizado Redes de Neuronas Artificiales, Árboles y Reglas de Regresión y distintos algoritmos de generación de conjuntos de clasificadores (meta-heurísticos). Sin embargo, es importante señalar que los primeros modelos se generaron el día 21 de marzo con los datos disponibles hasta ese momento. Se llevó a cabo un estudio comparativo de los distintos algoritmos de regresión. Los algoritmos que mejores resultados obtuvieron fueron $M5$ (algoritmo de generación de árbol de regresión) [8], $M5rules$ (algoritmo de generación de reglas de regresión) [9] y $AdditiveRegression$ (algoritmo de generación de conjuntos de regresores) [10].

VI-C. Pronósticos para Panamá

Como ya se ha comentado, estos sistemas dependen mucho de la cantidad y calidad de los datos. Por esta razón, a pesar de que se obtiene modelos que se pueden considerar buenos porque describen los datos disponibles, llevar a cabo predicciones más allá de dos o tres días vista no se considera oportuno. Por ejemplo, se ha visto que durante los fines de semana el número de pruebas que se realiza en la mayoría de los países disminuye lo cual produce un repunte de casos los primeros días de la siguiente semana.

Para realizar el pronóstico de los próximos días se han utilizado dos variaciones del conjunto de datos. En el primer conjunto de datos se utilizan todos los datos disponibles en el fichero original ($Orig.$) en donde cada región/país tiene asociada una serie temporal y datos de localización. En el segundo conjunto de datos se han eliminado los atributos referentes a la provincia/estado y la región/país (Sin_pais). Para cada uno de los conjuntos de datos, en la Figura 10, se pueden ver el pronóstico obtenido y el algoritmo utilizado para la generación del modelo.

Como se puede apreciar en la Figura 10, existen una diferencia en cuanto a los pronósticos realizados por los modelos generados en función del conjunto de datos utilizados. Esto se debe a que al eliminar los atributos de provincia/estado y país/región, el algoritmo intenta generalizar sin datos específicos del país generando modelos muy distintos. Por otro lado, el modelo creado con el conjunto original de datos tiene en

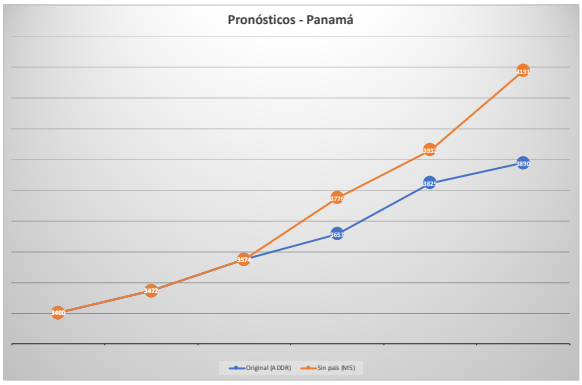

Fig. 10. Pronóstico de infectados acumulados para los próximos tres días

cuenta datos de los países y regiones para aplicar correcciones sobre los modelos lineales asociados a los árboles de regresión (ver Figura 11.

Fig. 11. Ejemplo de árbol de regresión con datos de Países/regiones

Es importante señalar que es necesario un enriquecimiento del conjuntos de datos, por ejemplo, con el cálculo de nuevos atributos y la fusión de otras fuentes de datos. Esto permitiría obtener modelos que, además de adecuarse muy bien a los datos que se poseen, sean capaces de llevar a cabo predicciones más acertadas.

VII. Conclusiones

Los modelos empleados hasta la fecha por el Minsa (Ministerio de Salud de Panamá) nos han mostrado proyecciones basadas en diferentes R0, al ser estos siempre mayores que uno nos llevan a escenarios no reales donde la mayoría de la población se infecta.

Este modelo epimediológico SIR donde los valores de $\beta(t)$ y $\gamma(t)$ son dependientes del tiempo muestran una estimación que se ajusta a los cambios y nos permite dar respuesta a cuando será el pico y que cantidad de infectados tendremos.

Igualmente hemos usado técnicas de análisis de datos para la estimación a corto tiempo basados en el comportamiento mundial de la enfermedad para tener un enfoque diferente.

Podemos concluir la importancia que tiene la uniformidad y calidad de los datos para mejorar las proyecciones. Y recomendamos que se realicen más test y sobre todo que se mantenga una continuidad en su cantidad independientemente si es fin de semana.

References

[1] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, vol. 115, no. 772, pp. 700–721, 1927.
[2] Y.-C. Chen, P.-E. Lu, and C.-S. Chang, "A time-dependent sir model for covid-19," *arXiv preprint arXiv:2003.00122*, 2020.
[3] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.
[4] (2020, Apr.) Help us better understand covid-19. Kaggle. [Online]. Available: https://www.kaggle.com/covid19
[5] (2020, Apr.) How to fight covid-19 with machine learning: 9 ways machine learning helps us fight the viral pandemic. Towards Data Science. [Online]. Available: https://towardsdatascience.com/fight-covid-19-with-machine-learning-1d1106192d84
[6] R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining," in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Springer-Verlag London, UK, 2000, pp. 29–39.
[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software," vol. 11, pp. 10–18, 2009.
[8] R. J. Quinlan, "Learning with continuous classes," in *5th Australian Joint Conference on Artificial Intelligence*. Singapore: World Scientific, 1992, pp. 343–348.
[9] G. Holmes, M. Hall, and E. Prank, "Generating rule sets from model trees," pp. 1–12, 1999.
[10] J. H. Friedman, "Stochastic gradient boosting," vol. 38, pp. 367–378, 2002.

(Este artículo no ha sido revisado por pares externos.)
(Artículo en constante actualización)
(Última actualización: 15/04/20)