ASTROLABIO
REVISTA DE CIENCIAS Y HUMANIDADES

UACM
Universidad Autónoma de la Ciudad de México
Nada humano me es ajeno

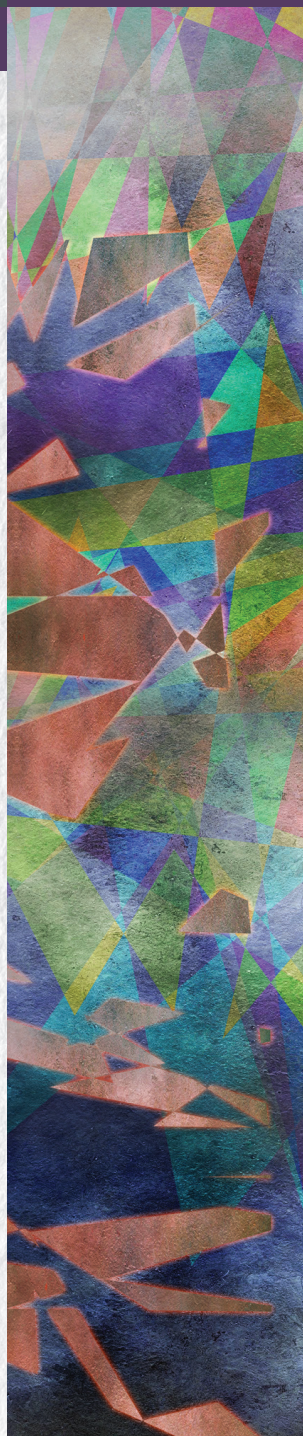
NÚMERO 12, INVIERNO 2023

ECLÍPTICA

Sección monográfica

Escenarios de propagación de COVID-19: modelación basada en agentes y clasificación supervisada

Alexis Nathán Rueda
Adán Fermín Castro Añorve
Araceli León Estrada
David Tusie Luna



www.uacm.edu.mx/astrolabio

Año 7, núm. 12, segundo semestre de 2023, es una publicación semestral de carácter académico editada por la Universidad Autónoma de la Ciudad de México, a través del Colegio de Ciencias y Humanidades. Calle Dr. García Diego núm. 168, col. Doctores, alcª. Cuauhtémoc, 06720, CDMX.
Editor responsable: Lidia Ivón Borja Aldave
ISSN 2594-231X. Reserva de Derechos al Uso Exclusivo 04-2018-110113192300-102.
Licitud de Título y Licitud de Contenido otorgados por la Comisión Calificadora de Publicaciones y Revistas Ilustradas de la Secretaría de Gobernación.

Escenarios de propagación de COVID-19: modelación basada en agentes y clasificación supervisada

Alexis Nathán Rueda
Adán Fermín Castro Añorve
Araceli León Estrada
David Tusie Luna

La evaluación de escenarios epidemiológicos se caracteriza por una elevada complejidad debido al acoplamiento de sistemas biológicos y sociales. En el contexto de la COVID-19, mundialmente hubo propuestas políticas de confinamiento a manera de intervenciones no farmacológicas con la finalidad de reducir la propagación de la enfermedad; sin embargo, la movilidad humana es un fenómeno de alta complejidad y su papel en la toma de decisiones en salud pública debe ser estudiado a profundidad. Con base en las características de la COVID-19 se diseñaron simulaciones computacionales de una epidemia usando un modelo basado en agentes. Se intervino sobre la movilidad de los agentes en un contexto epidémico, lo que permitió obtener escenarios comparables cuyos datos se analizaron con apoyo de algoritmos de clasificación supervisada. Se sugiere implementar nuevas herramientas y métodos para el estudio de la epidemiología y la movilidad humana en el ámbito de la salud pública.
Palabras clave: algoritmos de clasificación supervisada, modelación basada en agentes, COVID-19, esparcimiento de contagios, epidemias

COVID-19 spreading scenarios: agent-based modeling and supervised classification

Abstract

The assessment of epidemiological scenarios is characterized by a high complexity due to the coupling of biological and social systems. In the context of COVID-19, confinement policies were proposed worldwide as non-pharmacological interventions in order to reduce the spread of the disease; however, human mobility is a highly complex phenomenon and its role in decision-making in Public health must be studied in depth. Based on the characteristics of COVID-19, computational simulations of an epidemy were designed using an agent-based model. Agent mobility was intervened in an epidemic context, thus obtaining comparable scenarios whose data was analyzed with the support of supervised classification algorithms. The usefulness of implementing new tools and methods for the study of epidemiology and human mobility in the field of public health is suggested.

Keywords: supervised classification algorithms, agent-based modeling, COVID-19, spread of infections, epidemics

Introducción

Por *complejidad* entendemos el estudio de sistemas complejos adaptativos y, a su vez, a estos se les define como un conjunto de componentes que interactúan en forma directa o indirecta para modular su comportamiento (Hincapie-Palacio *et al*., 2013). La teoría de la complejidad estudia fenómenos como procesos históricos interconectados e incluye tanto sistemas sociales como ambientales, situando a los fenómenos epidémicos en una categoría intermedia, lo que conviene para su abordaje como sistemas complejos (Hincapie-Palacio *et al*., 2013; Rickles *et al*., 2007).

El presente trabajo es una aproximación al estudio de la dinámica de propagación poblacional de una enfermedad infecciosa desde la perspectiva de las ciencias de la complejidad. Se retoma la problemática de la COVID-19 para plantear escenarios epidémicos comparables desde la experimentación *in silico*, utilizando modelación basada en agentes (MBA) para luego analizar los resultados de dichas comparaciones con métodos de estadística computacional automatizada (*aprendizaje de máquina*). Para lograr lo anterior, es necesario revisar tres marcos conceptuales, los cuales se exponen brevemente a continuación.

Epidemiología de enfermedades infecciosas

Las enfermedades transmisibles son aquellas producidas por la interacción de agentes biológicos infecciosos específicos o sus productos tóxicos y se manifiestan por la trasmisión de estos (Módulos de Principios de Epidemiología para el Control de Enfermedades [MOPECE], 2001). Modelos conceptuales, como la *historia natural de la enfermedad* (HNE), permiten entender las distintas interacciones, agentes o circunstancias en las que se modifica el balance inicial entre el ser humano y su ambiente, favoreciendo la ocurrencia de infecciones (Rock *et al*., 2014). A través de la epidemiología puede cuantificarse el impacto social en salud pública que tienen las enfermedades infecciosas, utilizando indicadores (*e. g*., incidencia, prevalencia, mortalidad, etc.) que permitan medir y reconocer patrones en la dinámica de propagación de la enfermedad a escala poblacional. Al mismo tiempo, la epidemiología permite proponer modelos explicativos de índoles o escalas distintas que permitan evaluar la efectividad de una intervención sanitaria (Martcheva, 2015).

Modelación basada en agentes

El modelado basado en agentes (MBA) sigue un enfoque a partir del comportamiento de los agentes y de las interacciones entre ellos y con su medio ambiente, de donde emergen propiedades sistémicas generales (Wilensky y Rand, 2015), el uso de MBA permite capturar la complejidad del sistema a partir de la simulación de las acciones e interacciones de las entidades dentro del mismo (Badham *et al*., 2018). Autores diversos coinciden en que el MBA es idóneo para estudiar la complejidad multinivel de los sistemas sociales mediante la exploración de escenarios y la generación de información que facilite la explicación de los resultados sociales de las políticas públicas (Epstein, 2006). En este sentido, la dinámica que conduce a la propagación de una enfermedad infecciosa puede caracterizarse como un sistema complejo (Hunter y Kelleher, 2021) en el que ocurren procesos micro y macrosociales interdependientes, definidos tanto por parámetros poblacionales como por la heterogeneidad de los individuos (Coleman, 1990).

C. Algoritmos de clasificación

El aprendizaje de máquina o *machine learning* (ML) es una rama de la inteligencia artificial que emplea una amplia variedad de técnicas estadísticas, probabilísticas y de optimización que permiten a los sistemas computacionales *aprender* de grandes conjuntos de datos que son ruidosos y complejos, para detectar patrones, predecir tendencias o establecer esquemas de clasificación que, en la instancia de las metodologías estadísticas convencionales, serían difíciles de discernir (Cruz y Wishard, 2006; Mitchell, 1997; Cortes y Vapnik, 1995). Las técnicas de ML tienen una relevancia particular en el análisis de sistemas biológicos (*e. g*., la propagación de enfermedades infecciosas), pues muchos de estos sistemas son fundamentalmente no lineales y sus parámetros son condicionalmente dependientes.

El uso de ML en el ámbito de la salud tiene sus antecedentes hacia finales de la década de 1980 (Maclin *et al.*, 1991; Cicchetti, 1992) y actualmente forma parte de una tendencia creciente en el área clínica hacia una medicina predictiva personalizada (Weston y Hood, 2004). Este movimiento ha sido importante no solo para los pacientes y sus médicos sino para quienes, desde la administración de los sistemas de salud pública, tienen la responsabilidad de implementar políticas de prevención o tratamiento a gran escala (Weston y Hood, 2004).

En el contexto de la pandemia de COVID-19, la aplicación de ML y otras metodologías de inteligencia artificial han permitido mejoras significativas en el tratamiento, medicación, predicción, vigilancia y desarrollo de vacunas y fármacos. En todos los casos señalados, la inteligencia artificial no ha sido aplicada con la finalidad de reemplazar la intervención humana, sino de proveer una ampliación de perspectivas, mayor eficiencia y rapidez en el procesamiento e integración de múltiples tipos de información (proteómica, genómica, imagenológica, clínica, poblacional, etc.), para así consolidarse en un apoyo fundamental para hacer frente a la epidemia de SARS-COV-2 (Lalmuanawma *et al.*, 2020). A la fecha, en la gestión de los sistemas de salud, las técnicas de ML han supuesto una herramienta fundamental en la toma de decisiones para reducir la propagación de la COVID-19 (Vaishya *et al.*, 2020; Comito *et al.*, 2022).

Existe una amplia variedad de métodos computacionales para determinar patrones en los datos objeto de análisis; estos métodos, en su conjunto, se conocen como algoritmos de aprendizaje de máquina o de ML (Rusell y Norving, 1995) y, a su vez, pueden clasificarse en los siguientes paradigmas de aprendizaje: *aprendizaje supervisado*, *no supervisado* y *reforzado* (Soriano-Valdez *et al.*, 2021). En este trabajo se emplea únicamente el paradigma de aprendizaje supervisado, por constituir el enfoque metodológico de ML para el análisis de los resultados del MBA. En la evaluación de conjuntos de datos con este paradigma hay dos técnicas comúnmente utilizadas: *(a)* clasificación estadística y *(b)* análisis de regresión. Los algoritmos que se utilizan en este trabajo pertenecen a la técnica de clasificación estadística, en la que se constituyen los llamados *algoritmos de clasificación supervisada* (ACS).

Material y métodos

El modelo en NetLogo

Se utilizó *Epidemic-Lab* (Tusie, 2022; Tusie y Castro-Añorve, 2023), un modelo SEAIHRD (por sus siglas en inglés, que se exponen adelante) basado en agentes y desarrollado en NetLogo (Wilensky y Rand, 2015). *Epidemic-Lab* es un laboratorio virtual que permite visualizar, analizar y comparar escenarios de la dinámica espacial de propagación de una enfermedad infecciosa con características específicas, en poblaciones donde cada agente pueda tener los siguientes estados: susceptible (*S*), expuesto (*E*), infectado asintomático (*A*), infectado sintomático ambulatorio (*I*), hospitalizado (*H*, enfermo grave), recuperado (*R*) y difunto (*D*). Entre sus características relevantes para este trabajo, se encuentran la posibilidad de establecer la movilidad característica de los agentes (*i. e.*, la población modelada) en función de parámetros demográficos (*e. g.*, edad, género, etc.), establecer parámetros ajustables para

definir la HNE a modelar (tiempos medios, probabilidades, etc.) y aplicar períodos de confinamiento (Tusie, 2022). *Epidemic-Lab* constituye una herramienta útil en la formulación de políticas de control o gestión de una pandemia, así como en la prospección de epidemias con características de transmisión y prevención semejantes a las de COVID-19 (Tusie y Castro-Añorve, 2023).

Parámetros generales de entrada

En el modelo propuesto se representa tanto la evolución individual de la enfermedad como la dinámica de propagación poblacional, como fenómenos de contagio dependientes de la movilidad y confinamiento. Se tomaron parámetros probabilísticos a partir de la información disponible para caracterizar la HNE de la COVID-19, si bien se tomaron en cuenta los parámetros epidemiológicos de casos como el del crucero Diamond Princess (Rocklov *et al*., 2020), debe comprenderse que en epidemiología, no hay conjuntos universales de parámetros que logren describir por completo una enfermedad, y estos pueden variar significativamente según el lugar, el tiempo y otros factores contextuales (Gallo *et al.* 2020).

En el caso de COVID-19, en la etapa inicial de la pandemia (marzo de 2020), el número de infectados asintomáticos (capaces de transmitir la enfermedad) (Weitz *et al*., 2020), fue superior al número de enfermos y también debe destacarse que las manifestaciones de gravedad, incluyendo decesos, tendieron a aparecer solo en un limitado número de enfermos después de algunos días de evolución en función de eventos fisiopatológicos determinados en gran medida por condiciones individuales (*e. g*., edad y comorbilidades) (Ivorra *et al*., 2020). Así pues, en el modelo, el curso clínico de la enfermedad tiene un comportamiento probabilístico dinámico en cada individuo.

En el modelo, la variable de estado de los agentes corresponde a las etapas de la HNE en el contexto del mecanismo de transmisión; esto permitió representar a escala individual y poblacional las transiciones ocurridas en función de la exposición, infección y desenlace de la enfermedad como cambios de estado sujetos a una probabilidad.

El modelo también simula una movilidad típica de área urbana, en la que hay gente que vive en zonas céntricas, intermedias y periféricas, donde la tendencia es moverse de la periferia a *centros atractores de viajes* (que representan comercio, trabajo, educación, diversión, etc.) dispuestos en diversos lugares. Los agentes tienen un movimiento de tipo browniano, dirigido cada uno hacia determinado lugar habitual y también hay *saltos de Lévy* para simular desplazamientos largos, ya sea a centros atractores o hacia lugares particulares. Tanto los desplazamientos brownianos como los saltos de Lévy tienen una dirección y longitud estocástica, lo cual permite una riqueza en la variabilidad de movimientos origen-destino. En general, en la realidad, los patrones de movilidad humana a menudo son aleatorios y erráticos, como en el movimiento browniano, pero también ocurren por desplazamientos grandes y menos frecuentes, como los saltos de Lévy, por lo que estos dos tipos de movimiento permiten modelar matemáticamente la movilidad humana (Brockmann et al., 2006) y han sido implementados en otros modelos epidémicos estocásticos no lineales (El Koufi *et al*., 2022).

Datos de salida del modelo y su procesamiento

El modelo permite generar una *bitácora de simulaciones*, en la que se graba un renglón por simulación con los acumulados de veces que los agentes pasaron por cada estado, las duraciones en cada estado y valores generales. Así, por ejemplo, si un escenario epidémico consta de 80 simulaciones, entonces se genera una bitácora con 80 renglones.

A diferencia de otras técnicas de modelación, en el MBA los parámetros poblacionales son definidos *a priori* siguiendo una lógica experimental más que observacional (Squazzoni y Bianchi, 2023). En el modelo utilizado se generaron datos al simular escenarios epidémicos relacionados con la pandemia de COVID-19. Se aclara que estos datos no representan una instantánea directa de la situación epidemiológica de alguna entidad o población. En cambio, reflejan escenarios hipotéticos diseñados para evaluar el rendimiento de los ACS en un contexto de gestión de pandemias. La simulación de escenarios epidémicos permite crear un conjunto de datos controlado y diverso que abarca una variedad de condiciones y variables epidemiológicas relevantes. La estructura del conjunto de datos generado por el modelo se describe más adelante, en la sección titulada "Selección de parámetros de entrada y estructura de datos".

Algoritmos de clasificación supervisada

Los ACS utilizados en este trabajo son la máquina de soporte vectorial o de vectores (*Support Vector Machine* o SVM, por sus siglas en inglés) y el algoritmo de los *k*–vecinos más cercanos (*k*–Nearest Neighbors o *k*–NN).

SVM

Las SVM utilizan un conjunto de datos etiquetados para el *entrenamiento* de su aprendizaje y así obtener un hiperplano para que construya la superficie de decisiones y pueda aprender de ella para categorizar nuevos datos de entrada (Betancourt, 2005). Así, para el caso bidimensional, dados dos parámetros o variables de interés, en cuya gráfica de dispersión (espacio de parámetros) puedan clasificarse regiones de acumulación de puntos (clústeres), un algoritmo SVM determina la función matemática, llamada *kernel* (Cortes y Vapnik, 1995), que maximice la separación entre clústeres y forme una *frontera de decisión* alrededor de ellos. Así, un kernel mapea los datos graficados a un espacio de parámetros con una dimensión mayor. Las líneas de separación de los clústeres son los hiperplanos y están determinados por los subconjuntos de puntos de las clases, llamados *vectores de soporte*. En el área biomédica, los SVM son de gran utilidad en problemas en que la relación entre parámetros sea no-lineal (Cruz y Wishard, 2006) y como predictores del progreso clínico en enfermedades (Soriano-Valdez *et al.*, 2021).

k–NN

En este algoritmo, a una muestra de los datos de entrada se asigna la clase más frecuente de entre sus *k* vecinos más cercanos en el espacio de parámetros, según una medida de similitud (Altman, 1992). En la etapa de entrenamiento del algoritmo, son almacenados los vectores característicos y las etiquetas de clase del conjunto de datos de entrenamiento. Para evaluar a qué clase o categoría pertenece un nuevo conjunto de datos de entrada no etiquetados, el algoritmo asigna un vector a cada observación de dicho conjunto en el espacio de parámetros y, posteriormente, calcula la distancia entre cada uno de los vectores almacenados y los nuevos, seleccionando los *k* ejemplos más cercanos. Las distancias grandes entre conjuntos de observaciones indican disimilitud, en tanto que distancias cortas indican similitud.

La elección de estos algoritmos para analizar los resultados del MBA SEAIHRD está fundamentada en que los escenarios epidémicos simulados pudieron ser etiquetados *a priori,* en términos del nivel de propagación de la enfermedad infecciosa. Además, por su relativa simplicidad, estos algoritmos suponen un enfoque metodológico adecuado a una primera evaluación de los datos de salida del modelo, utilizando herramientas de aprendizaje de máquina. Así mismo, la elección de dos distintos ACS se sostiene en que, para que *Epidemic-Lab* pueda ser utilizado con carácter predictivo, es importante evaluar el desempeño de una técnica de clasificación con respecto a otra y luego determinar cuál de ambas tiene la mejor precisión predictiva.

Selección de parámetros de entrada y estructura de datos

Cada una de las dinámicas de confinamiento simuladas se encuentran descritas a detalle en la sección de la hipótesis de trabajo y estas constituyen las clases en el espacio de parámetros sobre las que los ACS son implementados.

Cada bitácora contiene los parámetros epidemiológicos característicos de las dinámicas de propagación y confinamiento. De estos parámetros, fueron elegidos aquellos con relevancia en la gestión de una pandemia, es decir, aquellos cuya manipulación a través de la implementación de medidas de confinamiento tengan un impacto tanto en la dinámica de propagación de contagios como en la ocupación hospitalaria. Estos parámetros se describen en la Tabla A.1 del Apéndice.

Los datos generados a partir de las bitácoras de *Epidemic-Lab* están estructurados a manera de una tabla de tamaño $M \times N$, con $M = 11$; los de parámetros epidémicos descritos en la Tabla A.1 y el número total de experimentos con distribución por escenario se describe en la sección siguiente.

Análisis estadístico

Estadística descriptiva de los experimentos computacionales

Se realizaron un total de 280 experimentos computacionales ($N = 280$) para los cuatro escenarios propuestos (70 experimentos por cada escenario), cuyos datos están consig-

nados en la Tabla A.2 del Apéndice. Esto permitió la utilización de medidas de resumen para comparar escenarios, al asumir razonablemente su distribución normal. Se analizaron distintos parámetros epidemiológicos para distintos grados de inmovilización de la población. Para cada variable se calcularon medidas de tendencia central, así como medidas de dispersión. Se consideró la pertinencia del uso de medidas paramétricas o no paramétricas. Se realizaron pruebas de normalidad (Shaphiro-Wilk y D'Agostino) para los conjuntos de datos de cada escenario en función de los parámetros estudiados.

Postprocesamiento de los datos

Una vez determinados los espacios de parámetros bivariados de mayor interés epidemiológico, las observaciones de cada parámetro fueron *re-escaladas*, utilizando la técnica de *estandarización*, en la cual el promedio de la distribución estadística de los datos se centra en el valor 0, en un rango que va de −1 a 1. Este procedimiento, conocido como *preprocesamiento* de datos (Soriano-Valdez *et al*., 2021) permite remover los sesgos creados por la diferencia en magnitud de las observaciones y es necesario, puesto que los ACS utilizan métodos de cómputo que numéricamente comparan las distancias entre puntos, como parte del proceso de clasificación (Altman, 1992; Betancourt, 2005).

Implementación de los ACS

Los ACS implementados en los espacios de parámetros epidemiológicos de interés son los contenidos en el paquete *scikit-learn* (versión 0.23.1) de Python, para *k*-NN y SVM. En particular, para el ACS SVM, se utilizó el tipo de clasificador *C-Support Vector Classification*, cuya formulación matemática del algoritmo es *ad hoc* para el abordaje de problemas multiclase, en que las variables no son necesariamente linealmente separables (Pedregosa *et al*., 2011). Las etapas de implementación de los ACS se describen a continuación.

Como parámetros de entrada de los ACS, la variable respuesta para ambos fue de tipo categórico y sus clases corresponden a los cuatro escenarios epidémicos simulados, mismos que se encuentran descritos, a detalle, en la sección de la hipótesis de trabajo. Cada clase, dada por un nivel de propagación de la enfermedad, fue codificada numéricamente en el orden del menor al mayor nivel de propagación. Por su parte, tanto para el algoritmo de *k*-NN como para el de SVM, el conjunto de datos de entrada, x, estuvo constituido por vectores 2–dimensionales, donde n es el número de observaciones en el conjunto de entrenamiento, obtenido de la partición descrita en la siguiente sección. Cada dimensión del vector representa un parámetro epidémico de los descritos en la Tabla A.1 del Apéndice. Los pares de parámetros epidémicos a evaluar con los ACS fueron seleccionados de acuerdo con los criterios de impacto epidemiológico descritos en la sección titulada "Selección de parámetros de entrada y estructura de datos" de este trabajo. También para ambos ACS, las etiquetas de clase correspondientes a cada vector de entrada se asignaron al conjunto de datos entrada asociado a la variable respuesta. En la sección A.4 del Apéndice se encuentra el vínculo al repositorio donde se encuentra el código de la implementación de los ACS escrito en lenguaje Python.

Partición del conjunto de datos

El conjunto de datos estructurado fue particionado en dos subconjuntos: uno de *entrenamiento* y otro de *prueba* para evaluar el desempeño de clasificación de cada ACS. Para evitar el *sobre-aprendizaje* del algoritmo (Rodvold *et al.*, 2001) y optimizar su capacidad de generalización, los tamaños de los conjuntos de entrenamiento y de prueba se definieron como 2/3 y 1/3 del tamaño del conjunto de datos original, respectivamente.

Optimización de los ACS

Para optimizar el desempeño de los algoritmos SVM y *k*-NN en términos de los parámetros de entrada de cada uno, se usó la herramienta *GridSearchCV* de *scikit-learn* para evaluar y seleccionar de forma sistemática las combinaciones de parámetros de cada ACS que determinara el mejor desempeño, *i. e.*, la mayor precisión predictiva de cada clasificador. Para cada ACS se exploraron los siguientes parámetros:

SVM: el coeficiente de regularización, distribuido en el intervalo [0.1, 1000]; el coeficiente, distribuido en [0.0001, 1] y el kernel, con funciones lineal, polinomial, rbf y sigmoide.

k-NN: el número de vecinos, *k*, en el intervalo [1, 12], y la métrica (distancia) de comparación numérica con distancias euclidiana, de Manhattan y de Minkowski, como opciones.

Evaluación del desempeño de los ACS

El desempeño de clasificación de cada ACS se evaluó cuantitativamente comparando el número de etiquetas predichas para el conjunto de prueba con las etiquetas originalmente asignadas en ese mismo subconjunto de datos. El módulo *scikit-learn* permite computar, de forma nativa, la precisión predictiva de cada ACS.

Hipótesis de trabajo

Enfoque general

Se propuso la movilidad como variable independiente (incluyendo políticas de confinamiento), mientras que la propagación de la enfermedad como variable de respuesta. La movilidad social humana es un fenómeno complejo, de modo que se configuraron diferencias entre escenarios para dicha variable en términos de políticas de inmovilización, para identificar posibles efectos en la propagación de la enfermedad.

Se plantea evaluar mediante los ACS y la comparación de indicadores epidemiológicos de los escenarios posibles, cuáles de ellos presentan un menor impacto en salud pública. A continuación, se explica esto con mayor detalle.

Escenarios de inmovilización aleatoria *v.s.* inmovilización selectiva

La diferencia entre escenarios de inmovilización es *aleatoria* o *selectiva*, ésta dada en función de la fracción de agentes *infectados sintomáticos* (*i*) y de los períodos de confi-

namiento que se establecen como parámetros de entrada en el modelo. Cuando la inmovilización es *aleatoria*, solo el 25% de los *I* se confina; mientras que cuando es *selectiva*, se confina al 50% de ellos. Además, cuando se utiliza inmovilización selectiva, también la misma proporción (25%) del total de los agentes son confinados en función de su grupo de edad y de su factor de vulnerabilidad modelados. La población vulnerable de agentes fue modelada en *Epidemic-Lab* a partir de los datos disponibles sobre los efectos iniciales de la COVID-19 en diferentes poblaciones, sobre los cuales la Organización Mundial de la Salud (OMS) fundamentó su respuesta inicial ante la pandemia, a saber, edad, género, afecciones médicas preexistentes, etc. (OMS, 2020; Patel *et al.*, 2020).

Dinámica de confinamiento y niveles de propagación epidémica

La dinámica de confinamiento que se simuló con *Epidemic-Lab* fue: un período de confinamiento intenso de 45 días, seguido de un segundo período de confinamiento laxo de 180 días, al que le siguió un tercer período de confinamiento intenso de 30 días.

Llamamos *NvProp Bajo* o de bajo nivel de propagación al escenario con confinamiento e inmovilización selectiva. Por su parte, nombramos *NvProp Intermedio* al escenario con inmovilización selectiva sin confinamiento, *i. e.*, un escenario de nivel de propagación intermedia de la enfermedad. El escenario *NvProp Alto* se refiere al escenario con confinamiento e inmovilización aleatoria y *NvProp Muy Alto* al escenario sin confinamiento e inmovilización aleatoria.

Resultados

Estadística epidemiológica y selección de parámetros de relevancia

Diversos autores consideran importantes los análisis y modelos estadísticos, no solo como criterio de validación en la implementación de modelos de aprendizaje de máquina aplicados a problemas de clasificación en salud pública y epidemiología, sino también para la selección de datos y ponderación de variables, según las características de los datos y el objetivo de la clasificación (Pellegrini *et al.*, 2020; Sun *et al.*, 2018). Los resultados del análisis estadístico y su discusión, sobre la que descansa la selección de variables de relevancia en la propagación de la enfermedad y la validación de las técnicas de clasificación, son dispuestos en las secciones A.2 y A.3 del Apéndice, por ser prohibitivamente extensos para los propósitos que a este trabajo atañen, sin que ello degrade su importancia.

Los parámetros que se consideran de relevancia epidemiológica como función del parámetro de movilidad, *sDist*, son *Incidencia* y proporción de casos hospitalizados (*P-hosp*). La *Virulencia* es un parámetro propio del agente infeccioso (MOPECE, 2020) y ello explica que no exhiba diferencias estadísticamente significativas entre escenarios; sin embargo, se utiliza en la implementación de los ACS para explorar la capacidad de estos en la discriminación de escenarios.

Se utilizó la variable *sDist* para definir la distancia recorrida por un agente, lo que en el modelo representa el desplazamiento diario de las personas al salir y regresar de su domicilio. A nivel individual, un aumento en el desplazamiento implica una mayor

cantidad de interacciones sociales y ambientales, *ergo* un mayor riesgo de contagio durante una epidemia, mientras que, a nivel poblacional, el aumento en la movilidad se puede asociar con una mayor cifra de casos (*Incidencia* acumulada) de la enfermedad y de hospitalizaciones (*P-hosp*).

Clasificación

Optimización y evaluación del rendimiento de los ACS

INCIDENCIA COMO FUNCIÓN DE *sDIST*. La figura 1 muestra los diagramas de decisión resultantes de cada algoritmo, para *Incidencia vs. sDist*. En la figura 1(a), se muestra el diagrama de decisión generado por SVM, con parámetros de optimización C = 100, gamma = 0.1 y un kernel polinomial. La precisión de clasificación del algoritmo, obtenida de evaluar el rendimiento con el subconjunto de prueba, fue del 95%. Por su parte, la figura 1(b), muestra el diagrama de decisión generado por *k*-NN. Los parámetros de optimización son *k* = 6 vecinos y una métrica de Minkowski. La precisión, determinada también con el correspondiente conjunto de prueba, fue del 96%.

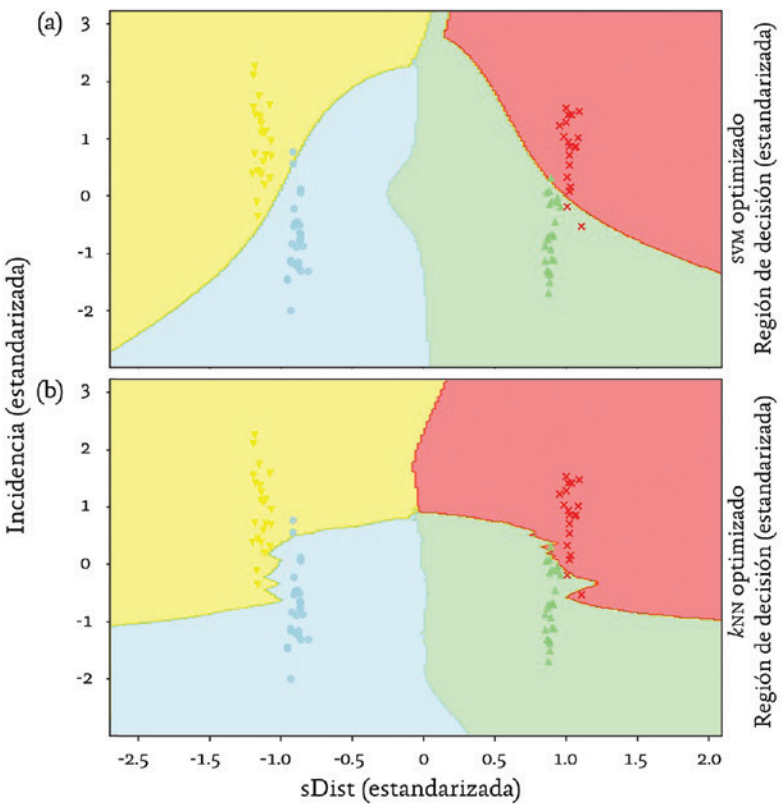


Figura 1. Diagramas de decisión resultantes de la implementación de los ACS en el espacio de parámetros *Incidencia v.s. sDist*.

P-HOSP COMO FUNCIÓN DE *sDIST*. Los diagramas de decisión para la clasificación del espacio de parámetros *P-hosp vs. sDist* se muestran como sigue: en la figura 2(a) para SVM, con C = 10, gamma = 0.1 y un kernel rbf, la precisión es del 95%. El diagrama de decisión generado por *k*-NN (*k* = 2 vecinos y una métrica euclidiana) para este mismo espacio de parámetros, se muestra en la fig. 2(b). La precisión obtenida fue 94%.

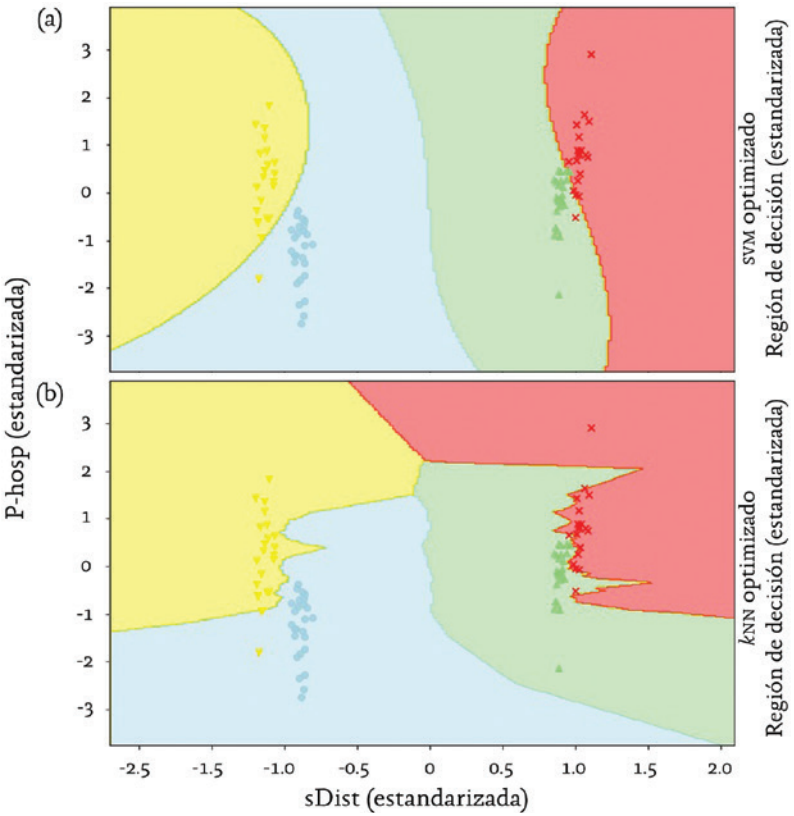


Figura 2. Diagramas de decisión resultantes de la implementación de los ACS en el espacio de parámetros *P-hosp vs. sDist*.

DISCUSIÓN. Si bien para ambos ACS la precisión de clasificación es numéricamente equivalente (~95%), las diferencias geométricas entre sus diagramas de decisión en los distintos espacios de parámetros indican que su selección para efectos de optimización es fuertemente dependiente de los datos. Además, como se observa en las figuras 1 y 2, valores numéricamente equivalentes en la precisión no implican equivalencia ni similitud geométrica de las regiones de decisión que se obtienen con uno y otro ACS, aun para un mismo espacio de parámetros, lo cual sugiere que la información predictiva que se obtenga de cada uno puede ser también distinta.

En el plano de clasificación formado por *Incidencia* como función de *sDist* se distinguen las cuatro regiones correspondientes a cada escenario. En la técnica de SMV puede advertirse cómo la delimitación entre escenarios ocurre principalmente a lo largo de

sDist, sin embargo, el escenario de propagación intermedia (*NvProp Intermedio*) puede tomar cualquier valor en el dominio de la *Incidencia*. Esto último no sucede con la técnica de *k*-NN, en la cual los escenarios *NvProp Bajo* y *NvProp Alto* ocupan la porción inferior del plano, lo que permite inferir argumentos para clasificar estos dos escenarios como los de menor propagación, dada su correspondencia con el eje de la *Incidencia*.

En el caso del plano de clasificación formado por *P-hosp* contra *sDist*, se advierte el mismo patrón observado en *Incidencia*, siendo incluso más clara la diferencia de escenarios en *NvProp Bajo* en comparación con *NvProp Alto* respecto a su relación de correspondencia con el eje de *P-hosp*.

La *Incidencia* acumulada es una medida de frecuencia de la enfermedad, que refleja el impacto social que tiene esta, como función del riesgo de enfermar. Esto, a su vez, permite estimar el grado de afectación de las actividades esenciales por días laborales perdidos por incapacidad o incluso los costos de la atención a la salud. Por su parte, *P-hosp* es un indicador del impacto social de la enfermedad porque refleja la carga asistencial del sistema de salud a nivel hospitalario (Bravo-García y Magis-Rodríguez, 2020).

VIRULENCIA COMO FUNCIÓN DE *P-HOSP*. Finalmente, para el espacio de parámetros de Virulencia como función de *P-HOSP*, la proporción de hospitalizaciones, se presenta el diagrama de decisión generados por SVM [figura 3(a)] con C = 1000, gamma = 0.1 y una función lineal como kernel, que determinaron 81% como la precisión mejor lograda. Por su parte, la precisión del algoritmo *k*-NN para este mismo espacio de parámetros, fue del 80%, con *k* = 1 vecinos y una métrica de Minkowski. El diagrama de decisión correspondiente se muestra en la figura 3(b).

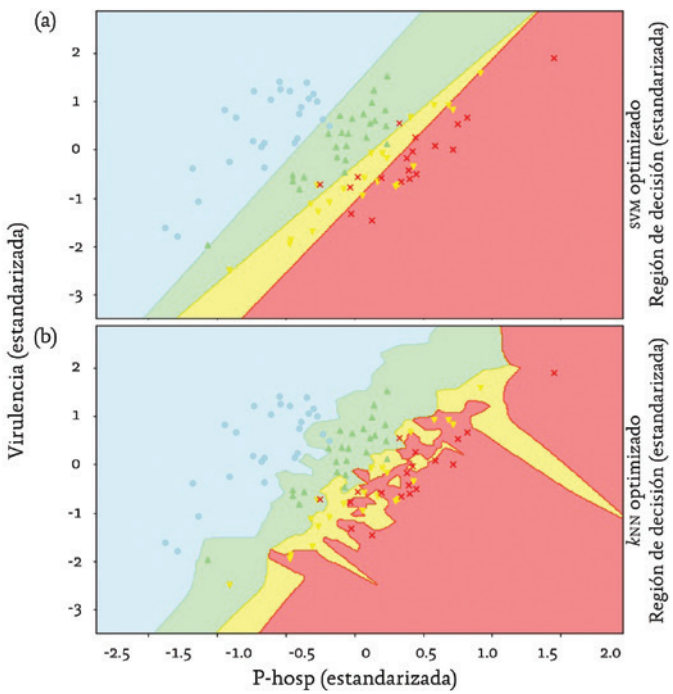


Figura 3. Diagramas de decisión resultantes de la implementación de los ACS en el espacio de parámetros de *Virulencia vs. P-hosp*.

Discusión. Se utilizó la *Virulencia* para explorar la capacidad de discriminación del ACS en un parámetro, de antemano identificado con variabilidad no significativa entre escenarios. El plano de decisión resultante de SVM [fig. 3(a)] exhibe geometrías bien delimitadas a lo largo del eje de *P-hosp*, sin embargo, todos los *Escenarios* abarcan el dominio del eje ordenado formando una diagonal, lo que significa que es posible encontrar cualquier valor de *Virulencia* para cada uno de ellos.

En el plano de decisión formado con *k*-NN [fig. 3(b)], si bien se conserva una imagen diagonal de orden y aspecto semejante al caso anterior, la superposición de *Escenarios* produce regiones altamente irregulares y discontinuas. Lo anterior ilustra la limitación de las técnicas de ACS para distinguir entre conjuntos de datos previamente reconocidos como indistintos. Adicionalmente, que la mayor precisión lograda en la optimización de ambos ACS no sea mayor al 81%, indica la pobreza predictiva de los ACS para casos donde no hay diferencias significativas entre escenarios. Lo anterior fundamenta la nula utilidad que tiene la implementación de los ACS para la clasificación de espacios bivariados con este tipo de parámetros y, además, sugiere la necesidad de la interpretación multidisciplinar de los resultados que se obtienen de la aplicación de técnicas de ML a problemas de salud pública.

Los resultados de este trabajo enfatizan que la selección de parámetros epidémicos relevantes desempeña un papel crítico en la implementación de ACS para la evaluación de la propagación de enfermedades infecciosas. La selección de estos parámetros se efectuó según su importancia desde una perspectiva epidemiológica, priorizando aquellos que tienen un impacto conocido en la propagación de contagios y la ocupación hospitalaria. Sin embargo, debe considerarse que la elección puede ser influenciada por sesgos o limitaciones en nuestro conocimiento actual de la enfermedad. Una alternativa es la reducción de dimensionalidad como enfoque para evaluar la relevancia de parámetros epidémicos (Agrebi y Larbi, 2020), mediante técnicas como *análisis de componentes principales* (*PCA*, por sus siglas en inglés) o *selección de características* (*feature selection*). Este enfoque supone la posibilidad de identificar de manera objetiva cuáles de ellas contribuyen significativamente a la variabilidad de los datos. No obstante, la reducción de dimensionalidad no reemplaza necesariamente la importancia de la experiencia epidemiológica en la selección de parámetros, pero puede servir como una herramienta adicional para validar o refinar suposiciones iniciales. Esto es especialmente relevante en situaciones en que la comprensión de la enfermedad está en constante evolución o cuando existen datos insuficientes para una evaluación epidemiológica exhaustiva.

Conclusiones

En el presente trabajo los resultados de la MBA permiten reconocer algunas relaciones entre la movilidad humana y las manifestaciones de la propagación de una enfermedad infecciosa en un contexto epidémico. Las políticas públicas de confinamiento se convirtieron en medidas de intervención no farmacológica aceptadas y utilizadas en todo el mundo en el contexto de la COVID-19; sin embargo, la mejor manera de implementar dicha intervención es un tema que aún debe estudiarse para comprenderse mejor, particularmente para intervenciones en eventos pandémicos futuros.

Por otro lado, puesto que el confinamiento de la población, el distanciamiento social y la reducción de la movilidad, como medidas integrales de rápida respuesta en salud pública frente a la propagación de enfermedades epidémicas, tienen profundas repercusiones socioeconómicas (OMS, 2020). Aquellos parámetros que cuantifiquen representativamente tales medidas constituyen una herramienta fundamental en lo que a toma de decisiones concierne. Algunas de las relaciones entre estos parámetros son particularmente difíciles de estudiar con métodos estadísticos convencionales.

Si bien es posible que la información con la que se opera habitualmente en la gestión de una epidemia se considere suficiente para los objetivos universalmente aceptados, y que asimismo, el estudio detallado de la movilidad con datos empíricos en medio de un escenario epidémico puede considerarse costoso o poco factible en términos materiales u organizativos, la necesidad de tomar decisiones críticas en la gestión de una epidemia justifica perseguir una mejor comprensión de los fenómenos que implica y su comportamiento a largo plazo.

Los resultados del trabajo permiten argumentar a favor de cómo la política de reducción de la movilidad tiene logros en la contención de la propagación de la enfermedad, siendo esto evidente en la reducción de la ocupación hospitalaria. En razón de lo anterior, este trabajo se presenta como una alternativa para el apoyo en el estudio de problemáticas en salud pública, ante retos epidemiológicos de la magnitud de la COVID-19.

Apéndice

El material suplementario a este trabajo se encuentra disponible en la siguiente liga para su consulta: https://bit.ly/45MQPZo

Referencias

Agrebi, S., y Larbi, A. (2020). Use of artificial intelligence in infectious diseases. En Barh, D. (Ed.), *Artificial intelligence in precision health* (pp. 415-438). Academic Press.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175-185. https://doi.org/10.2307/2685209

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

Badham, J., *et al.* (2018). Developing agent-based models of complex health behaviour. *Health & Place*, *54*, 170-177. https://doi.org/10.1016/j.healthplace.2018.08.022

Betancourt, G. A. (2005). Las máquinas de soporte vectorial (SVMs). *Scientia et technica*, *1*(27). https://revistas.utp.edu.co/index.php/revistaciencia/article/view/6895

Bravo-García, E., y Magis-Rodríguez, C. (2020). La respuesta mundial a la epidemia del COVID-19: los primeros tres meses. *Boletín sobre COVID-19, Salud Pública y Epidemiología*, *1*(1), 3-8.

Brockmann, D., *et al.* (2006). The scaling laws of human travel. *Nature*, *439* (7075), 462-465. https://doi.org/10.1038/nature04292

Cicchetti, D. V. (1992). Neural networks and diagnosis in the clinical laboratory: state of the art. *Clinical chemistry*, *38*(1), 9-10.

Coleman, J. S. (1990). *Foundations of Social Theory*. Harvard University Press.

Comito, C., y Pizzuti, C. (2022). Artificial intelligence for forecasting and diagnosing covid-19 pandemic: A focused review. *Artificial Intelligence in Medicine*, 102286. https://doi.org/10.1016/j.artmed.2022.102286

Cortes, C., y Vapnik, V. (1995). Support vector machine. *Machine learning*, *20*(3), 273-297. http://dx.doi.org/10.1007/BF00994018

Cruz, J. A., y Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, *2*, 59-77. https://doi.org/10.1177/117693510600200030

El Koufi, A. (2022). Nonlinear stochastic SIS epidemic model incorporating Lévy process. *Complexity*, 2022. https://doi.org/10.1155/2022/8093696

Epstein, J. M. (2012). *Generative social science: Studies in agent-based computational modeling*. Princeton University Press.

Gallo, L. G., *et al.* (2020). Ten epidemiological parameters of COVID-19: use of rapid literature review to inform predictive models during the pandemic. *Frontiers in Public Health*, *8*, 598547. https://doi.org/10.3389/fpubh.2020.598547

Hincapie-Palacio, D., y Ospina-Giraldo, J. F. (2013). La dinámica de la transmisión de la enfermedad según la teoría de la complejidad. *Revista de Salud Pública*, *15*, 655-665. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0124-00642013000500002&lng=en&tlng=es

Hunter, E., y Kelleher, J. D. (2021). Adapting an agent-based model of infectious disease spread in an Irish county to COVID-19. *Systems*, *9*(2), 41. https://doi.org/10.3390/systems9020041

Ivorra, B., *et al.* (2020). Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China. *Communications in nonlinear science and numerical simulation*, *88*, 105303. https://doi.org/10.1016/j.cnsns.2020.105303

Lalmuanawma, S., Hussain, J., y Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for COVID-19 (SARS-CoV-2) pandemic: A review. *Chaos, Solitons and Fractals*, *139*, 110059. https://doi.org/10.1016/j.chaos.2020.110059

Maclin, P. S., *et al.* (1991). Using neural networks to diagnose cancer. *Journal of Medical Systems*, *15*(1), 11-19. https://doi.org/10.1007/BF00993877

Martcheva, M. (2015). *An introduction to mathematical epidemiology* (Vol. 61, pp. 9-31). New York: Springer.

Mitchell, T. (1997). *Machine Learning*, New York: McGrawHill.

Módulos de Principios de Epidemiología para el Control de Enfermedades Transmisibles [MOPECE] (2001). Vigilancia en salud pública.

Organización Mundial de la Salud [OMS] (2020). Strategic preparedness and response plan for the new coronavirus. https://bit.ly/3M8UauX

Patel, A., y Jernigan, D. B. (2020). Initial public health response and interim clinical guidance for the 2019 novel coronavirus outbreak—United States, December 31, 2019–February 4, 2020. *American Journal of Transplantation*, *20*(3), 889-895. https://doi.org/10.1111/ajt.15805

Pedregosa, F., *et al.* (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Pellegrini, M., Ballerini Lippi, M., y Frumento, P. (2020). Machine learning in epidemiology: a review of algorithms and their applications to surveillance systems. *Epidemiology Biostatistics and Public Health*, *17*(2), e13648.

Rickles, D., Hawe, P., y Shiell, A. (2007). A simple guide to chaos and complexity. *Journal of Epidemiology and Community Health*, *61*(11), 933-937. http://dx.doi.org/10.1136/jech.2006.054254

Rock, K., *et al.* (2014). Dynamics of infectious diseases. *Reports on Progress in Physics*, *77*(2), 026602. DOI: 10.1088/0034-4885/77/2/026602

Rocklöv, J., Sjödin, H., y Wilder-Smith, A. (2020). COVID-19 outbreak on the Diamond Princess cruise ship: estimating the epidemic potential and effectiveness of public health countermeasures. *Journal of travel medicine*, *27* (3), https://doi.org/10.1093/jtm/taaa030

Rodvold, D. M., *et al.* (2001). Introduction to artificial neural networks for physicians: taking the lid off the black box. *The Prostate*, *46* (1), 39-44.

Rusell, S., y Norving, P. (1995). *Artificial Intelligence A Modern Approach*. Prentice-Hall.

Soriano-Valdez, D., *et al.* (2021). The basics of data, big data, and machine learning in clinical practice. *Clinical Rheumatology*, *40*(1), 11-23. https://doi.org/10.1007/s10067-020-05196-z

Squazzoni, F., y Bianchi, F. (2023). Exploring Interventions on Social Outcomes with In Silico, Agent-Based Experiments. En Damonte, A., Negri, F. (Eds.) *Causality in Policy Studies: a Pluralist Toolbox* (pp. 217-234). Springer, Cham.

Sun, Y., Wong, A. K., y Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence, 23* (04). https://doi.org/10.1142/S0218001409007326

Tusie, D. (2022). *Laboratorio epidemiológico para analizar la propagación de enfermedades transmisibles tipo COVID-19: una perspectiva desde el modelado basado en agentes.* Tesis de Maestría en Ciencias de la Complejidad. México: Universidad de Autónoma de la Ciudad de México. https://bit.ly/467df7k

Tusie, D., y Castro-Añorve, A. F. (2023). Modelo matemático de una epidemia o la práctica de la interdisciplina al servicio de la salud. *Astrolabio. Revista de Ciencias y Humanidades*, *10*, 22-34.

Vaishya, R., *et al.* (2020). Artificial Intelligence (AI) applications for COVID-19 pandemic. *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, *14*(4), 337-339. https://doi.org/10.1016/j.dsx.2020.04.012

Weitz, J. S., *et al.* (2020). Modeling shield immunity to reduce COVID-19 epidemic spread. *Nature Medicine*, *26*, 849-854. https://doi.org/10.1038/s41591-020-0895-3

Weston, A. D., y Hood, L. (2004). Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *Journal of Proteome Research*, *3*(2), 179-196. https://doi.org/10.1021/pr0499693

Wilensky, U., y Rand, W. (2015). *An introduction to Agent-Based Modeling. Modeling natural, social, and engineered complex systems with NetLogo*. MIT Press.