



CIENCIAS NATURALES Y EXACTAS

Artículo original de investigación

Adquisición de conocimiento sobre la letalidad de la COVID-19 mediante técnicas de inteligencia artificial

María Matilde García Lorenzo^{1*} <http://orcid.org/0000-0002-1663-5794>

Yanela Rodríguez² <https://orcid.org/0000-0003-2232-2265>

Alejandro Ramón-Hernández¹ <https://orcid.org/0000-0003-0749-5977>

Beatriz Bello-García¹ <https://orcid.org/0000-0002-3943-5736>

Yaima Filiberto⁴ <https://orcid.org/0000-0003-2279-2953>

Alejandro Rosete^{3,4} <http://orcid.org/0000-0002-4579-3556>

Yaile Caballero^{2,4} <http://orcid.org/0000-0002-6725-5812>

Rafael Bello^{1,4} <http://orcid.org/0000-0001-5567-2638>

¹ Universidad Central Marta Abreu; Villa Clara, Cuba

² Universidad de Camagüey Ignacio Agramante; Camagüey, Cuba

³ Universidad Tecnológica de La Habana José Antonio Echeverría (CUJAE); La Habana, Cuba

⁴ Miembro de la Academia de Ciencias de Cuba, La Habana, Cuba

* Autor para la correspondencia: mmgarcia@uclv.edu.cu

RESUMEN

Palabras clave

COVID-19; inteligencia artificial; aprendizaje automático; selección de rasgos; inducción de reglas

Introducción. La difusión de la COVID-19 en el mundo ha provocado una avalancha de investigaciones para enfrentarla y atenuar sus efectos, a lo cual han contribuido los estudios de inteligencia artificial. **Objetivos.** En este trabajo se muestra cómo pueden aplicarse técnicas de inteligencia artificial, en particular de ciencias de datos y aprendizaje automático, para obtener conocimiento útil en la predicción de la letalidad de la COVID-19. **Métodos.** Se analizaron los datos disponibles de pacientes en México hasta el 20 de abril, incluyendo 16 rasgos (físicos y clínicos) sobre cerca de 9000 casos positivos (más de 700 fallecidos), con el foco en identificar patrones que predigan un desarrollo fatal de la enfermedad. Se emplearon técnicas de preparación y visualización de datos, selección de rasgos e inducción de reglas empleando el algoritmo J48, las redes neuronales y la teoría de los conjuntos aproximados. **Resultados.** Los patrones encontrados por las diferentes vías empleadas coinciden en la relación fuerte entre varios rasgos de los pacientes en casos de letalidad, entre los cuales se destacan la edad, la obesidad, la hipertensión, la inmunosupresión, la diabetes, los problemas renales y cardíacos. Los resultados pueden ayudar a mejorar la comprensión de la enfermedad y muestran las capacidades de las técnicas de inteligencia artificial para analizar datos desde diferentes perspectivas, como apoyo al trabajo médico.



Knowledge acquisition about the lethality of COVID-19 by using artificial intelligence

ABSTRACT

Keywords

COVID-19; artificial intelligence; machine learning; feature selection; rule induction

Introduction. The spread of COVID-19 in the world has brought about much research to face it and alleviate its effects. The artificial intelligence community has actively participated in several dimensions of this effort. **Objective.** This paper shows how to apply Artificial Intelligence techniques, particularly Data Science and Machine Learning, in order to gain valuable knowledge to predict COVID-19 lethality. **Methods.** We analyzed the available data concerning Mexican patients until April 20, including 16 features (physical and clinical) of about 9000 positive cases (more than 700 deaths), focusing on identifying patterns to predict a fatal course of the disease. Several techniques were used for data preparation and visualization, feature selection and rule induction by using J48 algorithm, neural networks, and Rough Sets. **Results.** Patterns discovered through different ways coincide with the strong relationship among several features with respect to the lethality of COVID-19, in particular, age, obesity, hypertension, immunosuppression, diabetes, and renal and cardiac problems. Results permit a better understanding of the disease and they show the potential value of using Artificial Intelligence to do a multi-perspective analysis of clinical data to support medical work.

INTRODUCCIÓN

La COVID-19 es una enfermedad nueva cuyo surgimiento tuvo lugar en China a fines del año 2019. En la actualidad se ha extendido a prácticamente todo el mundo. Hasta mayo del 2020 se reconocía su presencia en 184 países, se habían reportado más de 4 millones de diagnósticos positivos de la enfermedad y decenas de miles de muertes.

Por su novedad, el conocimiento que se tiene es relativamente escaso desde muchos aspectos, aunque abundan las investigaciones en numerosos países. Estas incluyen estudios de su posible comportamiento, medios de diagnósticos y posibles tratamientos, y el desarrollo de las vacunas. Dada su relativa alta letalidad (aproximadamente del 7 %) comparada con otras generadas por virus similares, una de las temáticas de interés científico son las condiciones que pueden llevar a que peligre la vida de un paciente con la enfermedad. La inteligencia artificial es considerada una de las herramientas con gran potencialidad para enfrentar este reto.⁽¹⁾

Un ejemplo de estos trabajos es el presentado por Li Yan *et al.*,⁽²⁾ en el cual se desarrolla un modelo para clasificar la severidad y predecir una tendencia de desarrollo fatal de la enfermedad. En él se estudiaron varios centenares de casos, descritos inicialmente con más de 300 atributos, y se aplicaron técnicas de inteligencia artificial, en particular métodos de aprendizaje automático. En ese trabajo se llegaron a identificar tres atributos relevantes y se formularon reglas para predecir un desenlace negativo para un paciente. Los atributos identificados como relevantes fueron:

- LDH: lactato deshidrogenasa.

- L%: % de linfocitos.
- Hs-CRP: proteína C-reactiva de alta sensibilidad.

Las reglas descubiertas fueron las siguientes (SI indica fallecimiento; NO indica supervivencia):

- R1: si LDH ≥ 365 , entonces SI.
- R2: si LDH < 365 y hs-CRP $< 41,2$, entonces NO.
- R3: si LDH < 365 y hs-CRP $\geq 41,2$ y L% $> 14,7$, entonces NO.
- R4: si LDH < 365 y hs-CRP $\geq 41,2$ y L% $\leq 14,7$, entonces SI.

El descubrimiento de conocimiento en un dominio de aplicación a partir de los datos existentes en él es una práctica muy actual en estos momentos debido a los altos volúmenes de información que se genera y almacena. Se han desarrollado numerosas técnicas para el descubrimiento de conocimiento, desde la perspectiva de disciplinas como la estadística y la inteligencia artificial (IA). Los términos *aprendizaje automático*, *minería de datos* (y de textos) y *ciencia de datos* resultan familiares en la actualidad.⁽³⁾ El conocimiento generado a partir de la aplicación de estas técnicas puede usarse en la construcción de sistemas inteligentes, y también puede utilizarse para incrementar el conocimiento existente sobre el dominio de aplicación. Este último aspecto es especialmente útil en dominios donde el conocimiento todavía es limitado, como es el caso de la epidemia de COVID-19.

Entre las técnicas para el descubrimiento de conocimiento, existen algunas que permiten obtener un conocimiento comprensible por los expertos del dominio; por ejemplo, las reglas causales como las mostradas antes. Otros métodos

han demostrado ser muy eficaces para el descubrimiento de conocimiento y la construcción de sistemas inteligentes, pero el conocimiento descubierto resulta no comprensible por los expertos del dominio, siendo el caso de métodos como los basados en redes neuronales artificiales, las máquinas de soporte vectorial (*support vector machine*) y más recientemente el aprendizaje profundo (*deep learning*),⁽⁴⁾ los cuales permiten crear sistemas inteligentes eficaces para realizar la clasificación o diagnóstico, pero no son útiles para descubrir un conocimiento que enriquezca el saber sobre el dominio de aplicación. Para superar esta limitación se ha desarrollado una nueva línea de trabajo en la IA, denominada IA *explicable* (*explainable AI, XAI*),⁽⁵⁾ la cual permite desarrollar métodos para mejorar la interpretabilidad del conocimiento descubierto.

El propósito de esta investigación ha sido gestionar la adquisición de conocimiento útil para el pronóstico de un posible desarrollo fatal de la COVID-19 utilizando técnicas de IA. Nótese que el propósito de la investigación es similar a la desarrollada por Li Yan *et al.*,⁽²⁾ pero los rasgos usados para describir los casos son diferentes: mientras en aquella se utilizan resultados de exámenes de laboratorio (lactato deshidrogenasa, % de linfocitos y proteína C-reactiva de alta sensibilidad), en la presente investigación se usan otros tipos de atributos, como la edad, el sexo, si el paciente es obeso o no, las enfermedades que padece (diabetes, asma, hipertensión, etc.); también los datos utilizados son de origen distinto. Los datos utilizados han sido los publicados sobre la situación de la enfermedad en México (20). Se han aplicado técnicas de aprendizaje automático para identificar los rasgos más importantes para realizar el pronóstico y extraído conocimiento para esta problemática, el cual puede ser útil para confirmar patrones de comportamiento ya conocidos o eventualmente generar algunos novedosos. Los resultados son un primer paso en la generación del conocimiento; su validación solo es posible por parte de los especialistas en este campo.

MÉTODOS

Para desarrollar la adquisición de conocimiento se utilizaron los datos sobre el desarrollo de la COVID-19 en México, con fecha 20 de abril del 2020, la información de varios miles de pacientes incluyendo numerosos rasgos sobre estos, el diagnóstico realizado de la enfermedad y la fecha de muerte, en los casos de desarrollo fatal. Considerando el propósito de esta investigación los datos fueron depurados dejando los casos de pacientes que fueron diagnosticados con la enfermedad y se creó un atributo decisión donde se indica si el paciente falleció o no, denominado *letalidad*, cuyo dominio es {Letal= SI, Letal=NO}.

Como resultado se tiene una base de casos conformada por 8775 casos positivos a la COVID-19, con dos clases: 8060 pacientes con NO y 712 pacientes con valor SI, lo cual indica que los datos tienen un alto desbalance, es decir, de las dos clases de decisión una está mucho menos representada en los datos, en este caso los fallecidos. Está probado que el desbalance en los datos afecta los procesos de descubrimiento de conocimiento, lo cual fue tenido en cuenta para realizar el proceso de aprendizaje-descubrimiento.

Cada caso queda conformado por 16 atributos o rasgos predictores (tabla 1) y el rasgo de decisión o clase *letalidad*. De los 16 atributos hay un atributo numérico que corresponde a la edad del paciente y los 15 atributos restantes son discretos (nominales). Según el catálogo que describe los datos, cuando en el fichero aparece formato o fuente del tipo SI_NO cada celda puede contener uno de los siguientes valores (codificaciones): 1 equivale a SI, 2 codifica NO, 97 significa NO APLICA, 98 significa SE IGNORA, y 99 significa NO ESPECIFICADO.

Como primera acción en el estudio, se emplearon técnicas de visualización de información para mostrar la distribución de los datos según los casos analizados. Esta etapa de visualización es muy útil para entender los datos, se recomienda empezar por ella en los enfoques más actuales de ciencia de datos,⁽³²⁾ y en muchos casos permite enfocar los estudios más avanzados, empleando otras técnicas más complejas.

A partir de ahí, con esta base de casos se realizó el proceso de descubrimiento de conocimiento, en el cual se ejecutaron tres tareas principales:

1. Análisis de la consistencia de los datos.
2. Selección de rasgos importantes para predecir una evolución desfavorable del paciente.
3. Adquisición de conocimiento para la predicción.

El análisis de consistencia de los datos es importante porque tal y como se ha mostrado en algunos trabajos,^(7, 8) esta medida se relaciona con la calidad del conocimiento que luego se descubre al aplicar los métodos de aprendizaje automático; una alta consistencia de los datos permite una mayor eficacia de los métodos de aprendizaje.

El cálculo de la consistencia se basa en la medida calidad de la clasificación de la teoría de los conjuntos aproximados (del inglés *rough sets theory, RST*),^(7, 8, 9) la cual calcula el grado en que las clases están bien definidas, es decir, mide el grado de separación entre los objetos de las clases, en este problema las clases {Letal=SI, Letal=NO}. La esencia de la medida es que, si dos objetos son muy similares, inseparables y pertenecen a clases distintas, es decir, uno corresponde a un paciente fallecido y el otro es no fallecido, esto denota una inconsistencia en los datos, lo cual afectará el modelo de inferencia resultante del método de aprendizaje.

Tabla 1. Descripción de rasgos predictores

nº	Nombre del rasgo	Descripción	Formato o fuente
1	ORIGEN	La vigilancia centinela se realiza a través del sistema de unidades de salud monitoras de enfermedades respiratorias (USMER)	1 USMER 2 no USMER
2	SEXO	Identifica al sexo del paciente	1 mujer 2 hombre
3	TIPO_PACIENTE	Identifica el tipo de atención que recibió el paciente en la unidad. Se denomina como ambulatorio si regresó a su casa o se denomina como hospitalizado si fue ingresado a hospitalización	1 ambulatorio 2 hospitalizado
4	EDAD	Identifica la edad del paciente	Numérico
5	EMBARAZO	Identifica si la paciente está embarazada	SI_NO
6	DIABETES	Identifica si el paciente tiene un diagnóstico de diabetes	SI_NO
7	EPOC	Identifica si el paciente tiene un diagnóstico de EPOC	SI_NO
8	ASMA	Identifica si el paciente tiene un diagnóstico de asma	SI_NO
9	INMUSUPR	Identifica si el paciente presenta inmunosupresión	SI_NO
10	HYPERTENSION	Identifica si el paciente tiene un diagnóstico de hipertensión	SI_NO
11	OTRAS_COM	Identifica si el paciente tiene diagnóstico de otras enfermedades	SI_NO
12	CARDIOVASCULAR	Identifica si el paciente tiene un diagnóstico de enfermedades cardiovasculares	SI_NO
13	OBESIDAD	Identifica si el paciente tiene diagnóstico de obesidad	SI_NO
14	RENAL_CRÓNICA	Identifica si el paciente tiene diagnóstico de insuficiencia renal crónica	SI_NO
15	TABAQUISMO	Identifica si el paciente tiene hábito de tabaquismo	SI_NO
16	OTRO_CASO	Identifica si el paciente tuvo contacto con algún otro caso diagnosticado con SARS CoV-2	SI_NO

USMER: unidades de salud monitoras de enfermedades respiratorias.

Otra medida de RST utilizada para caracterizar los datos es la precisión de cada clase, que calcula en qué grado los casos de cada clase solo son inseparables de los objetos de esa clase.

El proceso de selección de rasgos consiste en determinar cuáles de los rasgos descritos en la tabla 1 tienen una mayor incidencia para realizar un pronóstico eficaz. Existen diferentes métodos para realizar este proceso. En este trabajo se utilizaron métodos de selección basados en búsqueda heurística y ordenamiento de los rasgos.

Para realizar el proceso de aprendizaje se han utilizado métodos de aprendizaje automático que permitan a los especialistas comprender el conocimiento descubierto. La forma de representación del conocimiento más comúnmente usada para construir sistemas basados en el conocimiento son las reglas; en este formalismo se construyen reglas de la forma:

si *Antecedente* entonces *Decisión*

La premisa de la regla (*Antecedente*) es una condición que se debe cumplir para que se tenga la *Decisión* que aparece como consecuente. Un ejemplo de ellas son las cuatro reglas presentadas en la *Introducción*. En ocasiones, el antecedente puede indicar una causa, pero en el caso de los datos

médicos generalmente indican un síntoma una característica observable.

Un método frecuentemente empleado para el descubrimiento de reglas se basa en la construcción de árboles de decisión (el cual fue el empleado en el estudio ⁽²⁾ comentado en la *Introducción*). Un árbol de decisión está formado por nodos, que corresponden a rasgos del dominio de aplicación, desde los cuales salen ramas que se corresponden con los valores del dominio del rasgo; los nodos terminales del árbol son clases del dominio de aplicación. El nodo raíz del árbol corresponde al rasgo que mayor separabilidad logra, y así se van seleccionando los rasgos por cada rama del árbol. El algoritmo ID3 y su extensión C4.5, ⁽¹⁰⁾ basado en la medida de entropía para la selección de los rasgos, es uno de los más usados con este propósito; en este trabajo se empleó el algoritmo J48, una implementación que aparece de este método en la plataforma para el descubrimiento de conocimiento Weka. ⁽¹¹⁾

Otra forma de representación del conocimiento que preserva la interpretabilidad son los prototipos. Un prototipo es un caso que sirve de patrón o representante de un conjunto de los casos que están en la base de casos. Los métodos basados en prototipos usualmente forman grupos de casos

similares y para cada grupo se determina un prototipo; este puede ser uno de los elementos del grupo o ser construido a partir de los casos que forman el grupo. En este trabajo se ha utilizado el método de construcción de prototipos basado en la granulación de la base de casos usando una relación de similitud, en el mismo se forman clases de similitud de los elementos de la base de casos y para cada clase se construye un prototipo.

Es reconocida la eficacia de las redes neuronales artificiales para resolver diferentes problemas de aprendizaje, entre ellos los de clasificación.⁽¹²⁾ En particular, las de redes conocidas como redes multicapas (*multi-layer perceptron*, MLP) en la cuales la topología de la red está formada por una capa de entrada (donde aparece una neurona por cada rasgo de entrada), una capa de salida (donde las neuronas representan las clases), y una o más capas intermedias, denominadas capas ocultas. Entre estas neuronas existen enlaces con pesos asociados; el algoritmo de aprendizaje es el encargado de encontrar el conjunto de pesos para un problema de aprendizaje particular a partir de la base de casos. El problema de las redes neuronales artificiales, como el caso del MLP, es que son modelos no interpretables,⁽⁵⁾ llamados *cajas negras*, pues el modelo resultante del aprendizaje puede ser eficaz en hacer el pronóstico, pero los especialistas no conocen el por qué se logra ese resultado.

Como el propósito de este trabajo es adquirir conocimiento a partir de los datos, no resulta útil un modelo de conocimiento como este, de allí la necesidad de utilizar las técnicas de XAI.

⁽⁵⁾ Entre los primeros métodos considerados en el campo de la XAI están los orientados a lograr interpretabilidad de las redes neuronales; en esta investigación se utiliza el método propuesto en el trabajo de Bullock.⁽¹⁾ Este método realiza un procesamiento del modelo de red neuronal aprendido y construye una regla para cada clase; aquí interesa especialmente la regla que permite inferir la clase de los fallecidos. Así, el método utilizado en la investigación es construir un modelo de red neuronal tipo MLP, utilizando como método de aprendizaje el de propagación de los errores hacia atrás (*backpropagation*), luego aplicar el método propuesto por Augusta⁽¹³⁾ para extraer las reglas.

RESULTADOS

A continuación se presentan los resultados obtenidos en el análisis de los datos, comenzando con la visualización de la información más relevante, fundamentalmente relacionada con las Defunciones. La figura 1 muestra la cantidad de personas contagiadas por el virus y la cantidad de defunciones por COVID-19 según cada valor de cada edad. En este caso, solo se incluyeron los datos registrados hasta el 18 de abril. Esta figura muestra que, afortunadamente, la cantidad de defunciones es una proporción pequeña para casi todos los valores de edad. Sin embargo, se observa que en los valores altos de edad, ambas series tienden a acercarse.

La figura 2 muestra esta misma relación de una manera más reveladora. En este caso, se muestra el porcentaje de defunciones con respecto al total (la letalidad), para cada edad.

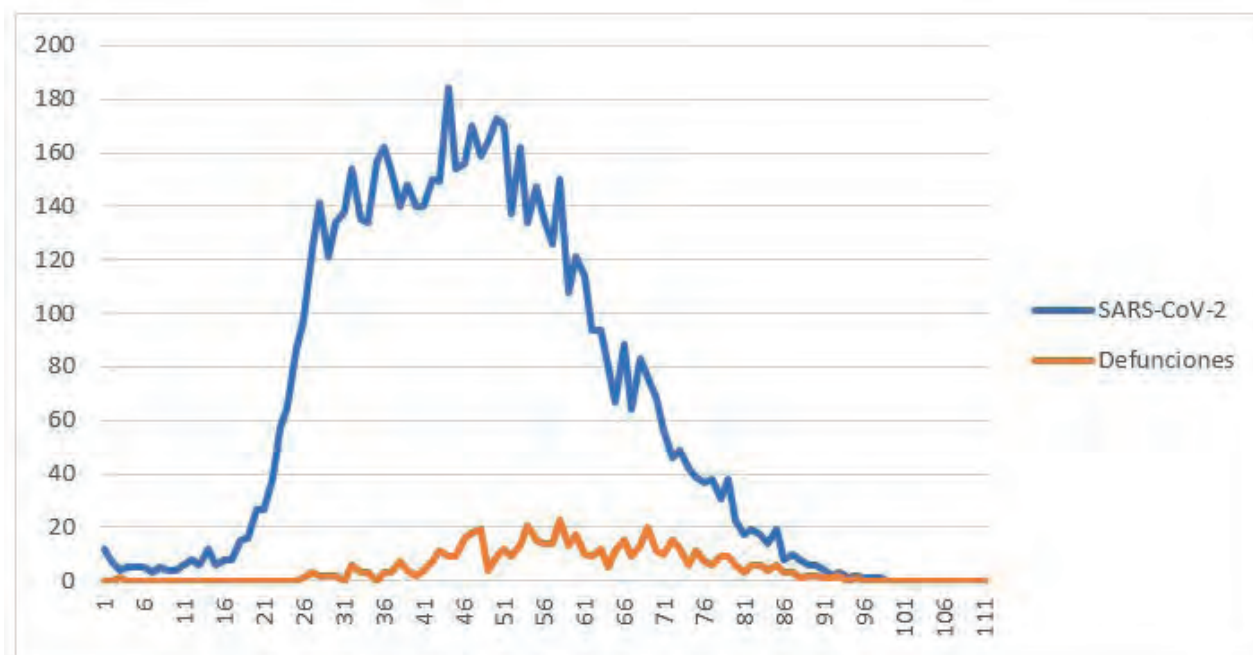


Fig. 1. Cantidad de personas contagiados y de defunciones según la edad.

Los valores reales aparecen en línea punteada. La línea continua es una versión suavizada obtenida a partir de promediar los valores adyacentes; es decir, la línea continua para la edad X muestra el promedio entre la letalidad de las edades X-1, X y X+1. En general, la letalidad hasta ese momento era del 7,94 %, pero se nota una clara tendencia a subir con la edad, siendo más de 0,1 (más del 10 %) a partir de los 55 años, y más del 20 % a partir de los 70 años, y mantiene una tendencia ascendente.

A partir de los gráficos mostrados en las figuras 1 y 2, se puede observar una tendencia a aumentar la letalidad con la edad. La influencia en la letalidad de varios de los otros factores descritos en la tabla 1 se muestra en la figura 3. En la figura 3 es notable cómo algunas de las características tienen una presencia proporcionalmente mucho mayor entre los fallecidos que entre todos los enfermos (con la excepción de la presencia del asma). Por ejemplo,

- los que padecen de problemas renales tienen una proporción 3,79 veces mayor entre los que fallecieron que entre todos los enfermos,
- la proporción de los que fueron diagnosticados por EPOC tiene una proporción 3,72 veces más entre los fallecidos,
- para los diagnosticados por neumonía es 2,8 veces mayores,
- 2,75 para INMUSUPR,

- 2,5 para problemas CARDIOVASCULARES,
- 2,27 para DIABETES y
- 2,1 para HIPERTENSIÓN.

Un aspecto colateral que vale la pena resaltar es que en la muestra analizada la letalidad es menor del 10 %. Esto implica que la proporción de ejemplo de fallecidos es alrededor de 1 por cada 9 ejemplos de los que sobreviven; es decir, los sobrevivientes (afortunadamente) constituyen una clase mayoritaria, lo cual es positivo. Sin embargo, esta desproporción o desbalance entre las clases debe atenderse con cuidado en los análisis de datos, porque tiende a influir en los conocimientos que se obtienen. ⁽³⁾

Esta descripción gráfica de la información permite poner en contexto el resto del estudio que se presenta. Cada una de las secciones siguientes describe el conocimiento sobre la COVID-19 que se puede obtener a partir del empleo de diferentes técnicas de aprendizaje automático.

Análisis de la consistencia de los datos

La medida *calidad de la aproximación* expresa la proporción de objetos que pueden ser correctamente clasificados en una clase dada, mientras que la medida *calidad de la clasificación* expresa la proporción de objetos que pueden clasificarse correctamente a partir de los datos disponibles. Utilizando la herramienta de análisis de datos basada en conjuntos apro-

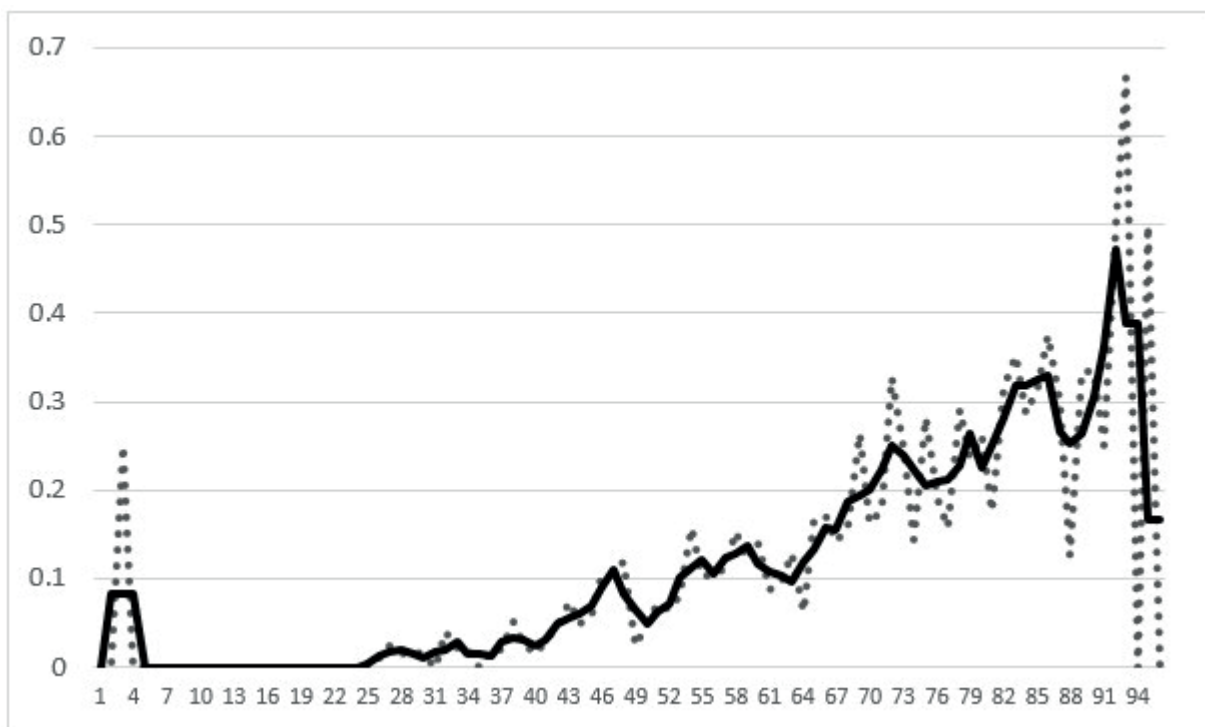


Fig. 2. Proporción de defunciones con respecto a los contagiados según la edad.

ximados,⁽⁹⁾ al aplicar las medidas para el análisis de los datos a la base de casos estudiada se obtienen los resultados siguientes:

- Calidad de la aproximación de la clase Letal=NO=0,76
- Calidad de la aproximación de la clase Letal=SI=0,37
- Consistencia de los datos (calidad de la clasificación)=0,73

Estos valores indican que la clase Letal= SI, además de ser una clase minoritaria en la base de casos, no tiene un valor alto de calidad; en este caso la causa es que los casos que tienen valor de decisión Letal= SI son similares a otros casos de la clase Letal=NO. Por su parte, el valor de la consistencia de los datos se puede considerar como satisfactorio (ni alto ni bajo) y esto incidirá posiblemente en que la eficacia de los métodos de aprendizaje automático.

Selección de los rasgos

Se usaron varios métodos de selección o pesado de rasgos, empleándose evaluación con validación cruzada de 10 particiones.

El objetivo de estos métodos es identificar aquellos atributos que tienen más peso a la hora de determinar si los datos son de una clase u otra (Letal= SI o Letal=NO), además de permitir una mejor visualización y comprensión de los datos.

Existen métodos, como los árboles de decisión, a los cuales no les afecta la presencia de atributos no significativos, ya que en el propio mecanismo de aprendizaje realizan una selección de atributos por su relevancia. Sin embargo, otros métodos no realizan este proceso, por lo que, si se realiza un

filtrado de atributos previo al aprendizaje, se puede mejorar de manera relevante su precisión, y al mismo tiempo se simplifican los modelos.

Para utilizar un método de selección de atributos lo primero es seleccionar el método de evaluación de atributos (*attribute evaluator*). Este método será el encargado de evaluar cada uno de los casos a los que se enfrente y dotar a cada atributo de un peso específico. El siguiente paso será elegir el método de búsqueda (search method), que será el encargado de generar el espacio de pruebas. Los resultados principales arrojados por los distintos métodos aparecen en la tabla 2.

Los métodos de búsqueda empleados fueron:

- Best-First, algoritmo genético (AG).⁽¹⁴⁾
- Un algoritmo clásico de pesado u ordenamiento de rasgos (RANK).

Como métodos de evaluación de los subconjuntos de atributos se utilizaron:

- CfsSubsetEval (evalúa el subconjunto de atributos, considerando la capacidad predictiva individual de cada rasgo y el grado de redundancia entre ellos).⁽¹⁵⁾
- Relief (evalúa el valor de un atributo muestreando repetidamente los casos y considerando el valor del atributo en casos cercanos de la misma clase y clases diferentes).^(16, 17, 18)
- SVM (evalúa el valor de un atributo usando el clasificador *máquina de soporte vectorial*).⁽¹⁹⁾
- PCA (análisis de componentes principales).⁽³⁾

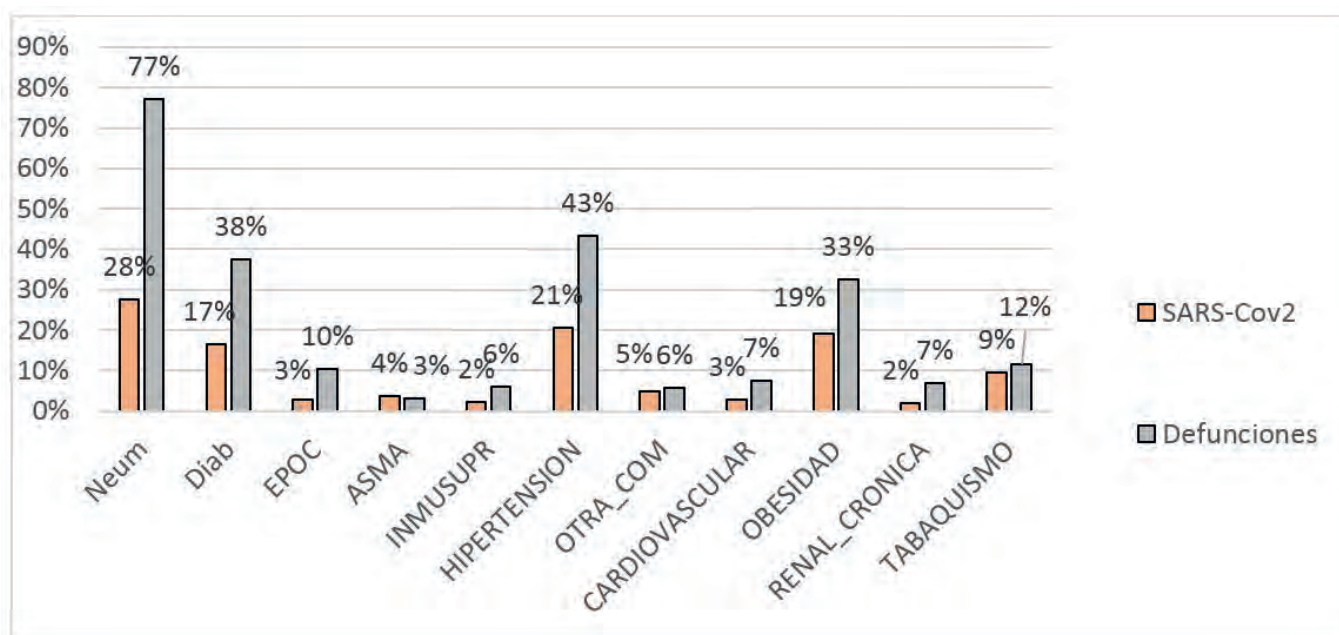


Fig. 3. Proporción de rasgos predictores entre los contagiados y entre los fallecidos.

Tabla 2. Resultados de los métodos de selección de rasgos

Método empleado	Rasgos seleccionados			
Best first con método de evaluación Cfssubseteval	Los rasgos TIPO_PACIENTE, EPOC, RENAL_CRONICA aparecen como los más relevantes en las 10 particiones utilizadas por el método de evaluación y en 8 de las particiones también aparece el rasgo OBESIDAD			
AG con método de evaluación Cfssubseteval	Los rasgos TIPO_PACIENTE, EPOC, RENAL_CRONICA aparecen como los más relevantes en las 10 particiones utilizadas por el método de evaluación y en 8 de las particiones también aparece el rasgo OBESIDAD			
Rank con método de evaluación Relief	Ordenamiento promedio resultante de la validación cruzada:			
	más relevantes		menos relevantes	
	1,1 ±0,3	TIPO_PACIENTE	9,2 ±1,66	EPOC
	1,9 ±0,3	EDAD	10 ±1,1	OTRA_COM
	3 ±0	OTRO_CASO	10,5 ±1,02	RENAL_CRONICA
	4,3 ±0,64	OBESIDAD	11,6 ±0,8	ASMA
	5,1 ±0,54	HIPERTENSIÓN	12,4 ±0,66	INMUSUPR
	5,9 ±0,54	DIABETES	13,7 ±0,46	TABAQUISMO
	7,8 ±1,17	CARDIOVASCULAR	15 ±0	SEXO
	8,5 ±3,01	ORIGEN	16 ±0	EMBARAZO
	El valor indicado es el lugar promedio alcanzado y su margen de error			
Rank con PCA	Para más del 41 % de la muestra sugiere que esta se puede representar por: 0,321inmusupr+0,321cardiovascular+0,32 hipertension+0,318asma+0,318EPOC... siendo los rasgos más importantes INMUSUPR, CARDIOVASCULAR, HIPERTENSION, ASMA y EPOC. Mientras, para el 28 % de la muestra considera que los rasgos que más influyen son SEXO, EMBARAZO y en menor medida TIPO_PACIENTE y EDAD, donde la ecuación: 0,69 sexo+0,689embarazo+0,19 tipo_paciente+0,096edad-0,06origen representa a esa muestra			
Rank con método de evaluación SVMatributeseval	Ordenamiento promedio resultante de la validación cruzada:			
	más relevantes		menos relevantes	
	3,4 ±2,2	INMUSUPR	7,7 ±3,9	OBESIDAD
	3,6 ±2,97	EPOC	9,1 ±3,59	HIPERTENSIÓN
	5,3 ±2,45	DIABETES	9,4 ±4	OTRA_COM
	6,7 ±4,31	ASMA	9,8 ±3,19	RENAL_CRONICA
	6,8 ±3,16	EMBARAZO	10,5 ±3,01	EDAD
	7 ±3,35	CARDIOVASCULAR	13,8 ±2,6	TIPO_PACIENTE
	7 ±3,03	TABAQUISMO	14,2 ±1,17	OTRO_CASO
	7,1 ±4,39	SEXO	14,6 ±1,56	ORIGEN
	El valor indicado es el lugar promedio alcanzado con un margen de error.			

Formalización del conocimiento para la predicción del peligro de fallecimiento

En esta sección se presenta el conocimiento adquirido formalizado como reglas de decisión y prototipos para la clase Letalidad= SI.

Reglas construidas usando árboles de decisión

J48 ha sido uno de los sistemas clasificadores más referenciados en la literatura, principalmente debido a su extrema robustez en un gran número de dominios como me-

dicina,^(20, 21, 22) agricultura,^(23, 24) electricidad,⁽²⁵⁾ educación,^(26, 27) ergonomía,⁽²⁸⁾ entre otros, y su bajo costo computacional, además de la facilidad con que pueden interpretarse sus resultados por cualquier usuario. En medicina se ha utilizado principalmente en el pronóstico de cáncer de mama. Este algoritmo trata eficazmente los valores desconocidos calculando la ganancia de información para los valores presentes y maneja los atributos continuos. Este algoritmo se encuentra dentro de los 10 más influyentes en minería de datos.⁽²⁹⁾

J48 puede ayudar no solo a hacer predicciones precisas a partir de los datos, sino también a explicar los patrones que contiene. Trata con atributos numéricos, valores faltantes, la poda, la estimación de las tasas de error, la complejidad de la inducción del árbol de decisión y la generación de reglas a partir de los árboles.

El parámetro más importante que deberemos tener en cuenta es el factor de confianza para la poda (*confidence level*), que influye en el tamaño y capacidad de predicción del árbol construido.

El árbol de decisión resultante de aplicar J48 (con factor de confianza 0,85 y un número mínimo de objetos de 10) permitió generar las seis reglas con mayor certeza CNF (valores entre 0,7 y 0,76, de un máximo posible de 1,0) para el caso de Letal=SI que se muestran en la tabla 3.

Los nodos con más peso son los que aparecen en la cima del árbol, pues a partir de una serie de procesos se seleccionan los atributos que mejor particionan el conjunto de datos, como puede apreciarse estos se corresponden con los obtenidos por los métodos de selección de atributos relevantes.

Reglas construidas usando una red neuronal artificial MLP y un método de XAI

Se desarrolló el proceso de construcción de una red neuronal tipo MLP el cual está compuesto por la topología siguiente:

- Una neurona en la capa de entrada por cada rasgo considerado,
- 20 neuronas en una capa oculta y
- 2 neuronas en la capa de salida, una por cada clase.

La red fue entrenada usando el método de propagación de los errores hacia atrás. Cuando se realizó el entrenamiento

utilizando el método clásico de ajuste de los pesos de este método la precisión alcanzada por el modelo resultante fue muy baja para la clase minoritaria (Letal=SI), que es precisamente la de mayor interés, en lo cual incide el desbalance de los datos. Por tanto se empleó una adaptación del método para el caso de datos desbalanceados, con lo que mejoró la eficacia del modelo construido y se obtuvo una precisión de 0,83 para la clase minoritaria.

Esta red entrenada constituye un modelo de conocimiento del dominio que permite la predicción para un paciente dado si tiene o no una tendencia a superar la enfermedad; sin embargo, el propósito de este trabajo no es la construcción de un sistema inteligente para realizar esta predicción, sino extraer conocimiento del dominio de aplicación. Por eso, usando el modelo de red neuronal construido se aplicó el método propuesto en Augasta⁽¹³⁾ para la extracción de reglas. La selección de este método se basó en dos criterios fundamentales: por una parte, el método incluye el análisis de qué rasgos son los más significativos para hacer la inferencia, lo cual significa que realiza una selección de rasgos pero directamente sobre el efecto de los rasgos en la inferencia y no como un método de pre-procesamiento como es usual en el aprendizaje automático, de hecho, numerosos métodos de extracción de reglas a partir de redes neuronales incluyen esta parte del método propuesto en Augasta⁽¹³⁾ por su efectividad; por otra parte, este método permite extraer muy pocas reglas, con lo cual la interpretabilidad de los resultados para los especialistas del dominio resulta mejor y, como se ha planteado antes, el propósito de esta investigación es facilitar la adquisición de conocimiento en este dominio, no construir un sistema de predicción inteligente. Este método determina los rasgos que son significativos para la predicción, eliminando los demás, y para ello determina cuánto se afecta la eficacia de la red

Tabla 3. Reglas obtenidas de los árboles de decisión

no.	Regla	cnf
r1	si EPOC=si and EDAD > 63 and OBESIDAD=si and DIABETES=si entonces letal=sí	0,71
r2	si EPOC=si and EDAD > 63 or HIPERTENSIÓN=si entonces letal=si	0,72
r3	si EPOC=no and RENAL_CRONICA=si and HIPERTENSIÓN=si or OBESIDAD=si entonces letal=si	0,7
r4	si EPOC=no and RENAL_CRONICA=no and HIPERTENSION=si and edad >58 entonces letal=si	0,75
r5	si EPOC=no and RENAL_CRONICA=no and TABAQUISMO=sí and EDAD<60 entonces letal=si	0,74
r6	si EPOC=no and RENAL_CRONICA=no and EDAD >64 and OBESIDAD=si entonces letal=si	0,76

construida si se quita de ella la neurona de entrada que representa el rasgo. Luego de este proceso los rasgos que se mantienen como significativos son {edad, embarazo, diabetes, inmunosuprimido, hipertensión, obesidad, renal-crónica, tabaquismo, otro caso}.

Este método determina una regla de decisión para cada clase. Como en este problema se tienen dos clases {Letal=SI, Letal=NO} se obtienen dos reglas. A los efectos de este trabajo resulta de interés la regla que determina el valor Letal=SI. La regla de decisión encontrada es la siguiente:

si (edad>57 and diabetes=si and hipertension=si and obesidad=si) entonces
letalidad=si

Al evaluar esa regla contra los casos de la base de casos correspondientes a pacientes fallecidos se tiene una precisión de 0,34. Este valor es coherente con el resultado del análisis de esta clase presentado en el epígrafe "Análisis de la consistencia de los datos" (calidad de la aproximación de la clase Letal=SI=0,37).

Conocimiento descubierto en forma de prototipos

La utilización de métodos de selección o generación de ejemplos (en el desarrollo de clasificadores, en particular, y en un proceso de aprendizaje, en general) puede mejorar la interpretabilidad ya que proporcionan información comprensible para el ser humano como salida. Esta información suele ser mucho más fácil de interpretar y comprender si está basada en un número reducido de ejemplos, denominados *prototipos*.

El método de construcción de prototipos utilizado en esta investigación se basa en construir una granulación del universo, es decir, determinar el conjunto de casos que son similares, y luego construir un prototipo para cada gránulo o conjunto de casos similares. Para construir la granulación se

utiliza el método propuesto por Filiberto *et al.* ^(30,31) para determinar la relación de similaridad más adecuada al problema. Este enfoque ha sido usado en otros métodos para la construcción de prototipos. ^(32,33)

Los algoritmos clásicos de construcción de prototipos tratan la clase positiva y la negativa en una forma simétrica, por lo que no están preparados para lidiar con el desbalance de clases (que, como se ha visto, está presente en este caso porque afortunadamente son mucho más los casos asociados a personas que sobreviven que los que fallecen). En este acápite se utiliza el método de construcción de prototipos según el método IMBNPBASIR SEL-CLASS, para la clasificación de conjuntos de datos desbalanceados. ^(34,35) La variante analizada consiste en la modificación de la medida calidad de la similitud, la cual sirve de base para el cálculo de los pesos usados en la selección de los conjuntos de prototipos. La novedad de la propuesta radica en el uso de la teoría de conjuntos aproximados, específicamente el uso de la medida calidad de la similaridad y los conceptos de clasificación basados en prototipos, para clasificar objetos con desbalance de clases.

Al realizar la clasificación basada en prototipos es necesario tener presente que estos son representantes de un grupo de casos, de modo que el cálculo de la precisión del conocimiento descubierto, es decir, del conjunto de prototipos, se realiza buscando para cada caso a cuál prototipo resulta más similar. Los prototipos seleccionados por el método IMBNPBASIR SEL-CLASS permiten realizar una clasificación con una exactitud de la clasificación general de 79,14. Teniendo en cuenta que la clase relevante en este estudio es la que denota una posible alta letalidad de la enfermedad seguidamente se presentan los mejores prototipos construidos para esa clase. Para evaluar la calidad de los prototipos se determina el conjunto de casos similares a él, y de ellos cuantos tienen el valor de decisión SI. Los prototipos obtenidos se muestran en la tabla 4.

Tabla 4. Prototipos obtenidos

No	Prototipo	cnf
p1	EDAD=60, DIABETES=si, EPOC=no, ASMA=si, INMUSUPR=no, HIPERTENSION=si, OTRA_COM=no, CARDIOVASCULAR=si, OBESIDAD=si, RENAL_CRONICA=no, TABAQUISMO=no, LETAL=si	0,63
p2	EDAD=76, DIABETES=si, EPOC=no, ASMA=no, INMUSUPR=si, HIPERTENSION=si, OTRA_COM=no, CARDIOVASCULAR=no, OBESIDAD=si, RENAL_CRONICA=si, TABAQUISMO=no, LETAL=si	1,00
p3	EDAD=65, DIABETES=si, EPOC=no, ASMA=no, INMUSUPR=si, HIPERTENSION=si, OTRA_COM=no, CARDIOVASCULAR=no, OBESIDAD=no, RENAL_CRONICA=no, TABAQUISMO=no, LETAL=si	0,61
p4	EDAD=65, DIABETES=si, EPOC=no, ASMA=no, INMUSUPR=no, HIPERTENSION=no, OTRA_COM=no, CARDIOVASCULAR=no, OBESIDAD=no, RENAL_CRONICA=si, TABAQUISMO=no, LETAL=si	0,57
p5	EDAD=60, DIABETES=si, EPOC=no, ASMA=no, INMUSUPR=no, HIPERTENSION=si, OTRA_COM=no, CARDIOVASCULAR=no, OBESIDAD=no, RENAL_CRONICA=no, TABAQUISMO=no, LETAL=si	0,55
p6	EDAD=72, DIABETES=si, EPOC=no, ASMA=no, INMUSUPR=si, HIPERTENSION=no, OTRA_COM=no, CARDIOVASCULAR=no, OBESIDAD=no, RENAL_CRONICA=si, TABAQUISMO=no, LETAL=si	0,55

Como un resultado de la construcción de la relación de similitud se obtiene también el peso de los rasgos del dominio de aplicación; para este problema se calcularon los pesos que se muestran en la tabla 5. Nótese la coincidencia, en general, con los resultados de la selección de rasgos presentada en la tabla 2.

Tabla 5. Peso de los rasgos según el método IRBASIR

No. orden	Rasgo	Peso	No. orden	Rasgo	Peso
1	EPOC	0,191323	8	OBESIDAD	0,044351
2	CARDIOVASCULAR	0,168401	9	ORIGEN	0,043986
3	RENAL_CRONICA	0,140204	10	TABAQUISMO	0,027925
4	DIABETES	0,111942	11	HIPERTENSION	0,021732
5	INMUSUPR	0,085462	12	ASMA	0,010395
6	EMBARAZO	0,081247	13	OTRO_CASO	0,008574
7	EDAD	0,055963	14	OTRA_COM	0,008495

Conclusiones

Resulta de interés que la selección de rasgos realizada usando diferentes métodos de selección produce resultados similares (tablas 2 y 5), lo que ratifica que algunos atributos inciden con mayor fuerza en la letalidad. En el caso del conocimiento inferido sobre la letalidad de la enfermedad, tanto las reglas como los prototipos representan patrones para la predicción del riesgo de muerte para el paciente, lo que permite identificar diferentes alternativas que pueden elevar el riesgo de muerte.

Los patrones encontrados por las diferentes vías empleadas coinciden en la relación fuerte entre varios rasgos de los pacientes con la letalidad de la enfermedad, entre los que destacan la edad, la obesidad, la hipertensión, la inmunosupresión, la diabetes, los problemas renales y los cardíacos.

Los resultados alcanzados en este trabajo, basados en el empleo de las técnicas de IA, permiten ratificar conocimiento existente sobre esta nueva enfermedad y también pudieran ofrecer a los especialistas algunas aristas novedosas sobre su letalidad.

REFERENCIAS BIBLIOGRÁFICAS

- Bullock J., Luccioni A., Hoffmann-Pham K., Nga-Lam C.S., Luen-go-Oroz, M.: MAPPING THE LANDSCAPE OF ARTIFICIAL INTELLIGENCE APPLICATIONS AGAINST COVID-19, United Nations Global Pulse - Durham University - Universite de Montreal - NYU Stern School of Business - World Health Organization. Available from <https://arXiv:2003.11336v1.2020>.
- Li Yan et al. Prediction of criticality in patients with severe COVID-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. 2020. preprint DOI: <https://doi.org/10.1101/2020.02.27.20028027>.

- Skiena S.S. The Data Science Design Manual. Springer. 2017.
- Skansi S. Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence. Springer Nature. 2018.
- Gunning D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.-Z. XAI—Explainable artificial intelligence. *Sci. Robot.* 2019;4.
- BC COVID-19 México 20 de abril 2020. Available from <https://www.gob.mx/salud/documentos/datos-abiertos-152127>
- Caballero, Y., Bello, R., et al. New Measures for Evaluating Decision Systems using Rough Set Theory: The Application in Seasonal Weather Forecasting. in Third International ICSC Symposium on Information Technologies in Environmental Engineering (ITEE'07). Carl von Ossietzky Universität Oldenburg. Alemania: Springer Verlag. 2007.
- Caballero, Y., Bello, R., et al. Knowledge Discovery using Rough Set Theory. In *Advances in Machine learning I Dedicated to the memory of Professor Ryszard S. Michalski in Series: Studies in Computational Intelligence*; Koronacki J., Ras Z.W., Wierzchon S.T.; Kacprzyk J (Eds), ISBN 978-3-642-05176-0. 2010; 262: 367-383.
- Bello-García, B. et al. Implementación de métodos para el pre-procesamiento de datos usando teoría de los conjuntos aproximados (RST) en Python. *Memoria de la Conferencia Internacional de Procesamiento de la Información (CIPI2019)*, Cuba. ISBN 978-959-312-372-3. 2019.
- Mitchell Tom M. *Machine Learning*. McGraw-Hill Science/Engineering/Math. 432 pages. ISBN: 0070428077. 1997.
- Herramienta de código abierto escrita en Java. Disponible bajo licencia pública GNU en <http://www.cs.waikato.ac.nz/~ml/weka/>.
- J. M. Zurada. *Introduction to Artificial Neural Systems*. West Publishing. 1992.
- Augasta, M.G., Kathirvalavakumar, T.: Reverse engineering the neural networks for rule extraction in classification problems. *Neural processing letters*. 2012; 35(2):131-150.
- David E. Goldberg. *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley.1989.
- M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning*. Hamilton, New Zealand.1998.
- Kenji Kira, Larry A. Rendell: A Practical Approach to Feature Selection. In: Ninth International Workshop on Machine Learning.1992: 249-256.
- Kononenko Igor (1994): Estimating Attributes: Analysis and Extensions of RELIEF. In: European Conference on Machine Learning, 171-182, 1994.
- Marko Robnik-Sikonja, Igor Kononenko: An adaptation of Relief for attribute estimation in regression. In: Fourteenth International Conference on Machine Learning.1997: 296-304.
- Guyon, J. Weston, S. Barnhill, V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*.2002; 46: 389-422.
- Hamsagayathri, P., & Sampath, P. Priority based decision tree classifier for breast cancer detection. *Advanced Computing and Communication Systems (ICACCS)*. 2017 4th International Conference. IEEE 2017: pp. 1–6. <https://doi.org/10.1109/ICACCS.2017.8014598>.
- Jhajharia, S., Verma, S., & Kumar, R. A cross-platform evaluation of various decision tree algorithms for prognostic analysis of breast cancer data. In *Inventive Computation Technologies (ICICT)*, International Conference. 2016; 3: pp. 1-7. IEEE. <https://doi.org/10.1109/INVENTIVE.2016.7830107>.

22. Muthuselvan, S., & Soma, K. (2016). Prediction of Breast Cancer Using Classification Rule Mining Techniques in Blood Test Datasets. En International Conference On Information Communication And Embedded System (ICICES 2016). IEEE. <https://doi.org/10.1109/ICICES.2016.7518932>.
23. Mishra, S., Paygude, P., Chaudhary, S., & Idade, S. Use of Data Mining in Crop Yield Prediction. En Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018). 2018. <https://doi.org/10.1109/ICISC.2018.8398908>.
24. Phadikar, S., & Goswami, J. Vegetation Indices Based Segmentation for Automatic Classification of Brown Spot and Blast Diseases of Rice. En 3rd Int'l Conf. on Recent Advances in Information Technology. 2016. IEEE. <https://doi.org/10.1109/RAIT.2016.7507917>.
25. Da Silva Pessoa, A. L., & Oleskovicz, M. Fault location in radial distribution systems based on decision trees and optimized allocation of power quality meters. En PowerTech, 2017 IEEE Manchester. 2017:1-6.
26. Pandey, A. K., & Rajpoot, D. S. A comparative study of classification techniques by utilizing WEKA. En Signal Processing and Communication (ICSC), 2016 International Conference. 2016: 219-224. IEEE. <https://doi.org/10.1109/ICSPCom.2016.7980579>.
27. Sivasakthi, M. (2017). Classification and Prediction based Data Mining Algorithms to Predict Students' Introductory programming Performance. 2017 International Conference on Inventive Computing and Informatics (ICICI). 2017: 346-350. <https://doi.org/10.1109/ICICI.2017.8365371>.
28. Rajesh, R., Maiti, J., & Reena, M. Decision Tree for Manual Material Handling Tasks Using WEKA. En P. K. Ray & J. Maiti (Eds.), Ergonomic Design of Products and Worksystems - 21st Century Perspectives of Asia. Singapore: Springer Singapore. 2018: 13-24. https://doi.org/10.1007/978-981-10-5457-0_2.
29. Bashir, U., & Chachoo, M.A. Performance Evaluation of J48 and Bayes Algorithms for Intrusion Detection System. International Journal of Network Security & Its Applications. 2017; 9:01-11.
30. Filiberto, Y., et al., Algoritmo para el aprendizaje de reglas de clasificación basado en la teoría de los conjuntos aproximados extendida. Dyna. 2011; 78(169): 62-70.
31. Filiberto, Y., et al. An analysis about the measure quality of similarity and its applications in machine learning. in Fourth International Workshop on Knowledge Discovery, Knowledge Management and Decision Support. 2013. Atlantis Press.
32. Bello-García, M., M.M. García-Lorenzo, and R. Bello. A method for building prototypes in the nearest prototype approach based on similarity relations for problems of function approximation. Lectures Notes on Computer Sciences. 2012; 7629.
33. Frías, M., et al. Prototypes selection based on similarity relations for classification problems in Engineering Applications - International Congress on Engineering (WEA2015). IEEE Press. Bogota, Colombia. 2015.
34. Rodríguez, Y., et al., An approach for class imbalanced data classification based on Rough Set and Nearest Prototype. , in 2nd International Symposium on Fuzzy and Rough Sets (ISFUROS 2017), Springer, Editor. Santa Clara, Cuba. 2017.
35. Rodríguez, Y., et al., Similar Prototype Methods for Class Imbalanced Data Classification, in Uncertainty Management with Fuzzy and Rough Sets. Springer. 2019: 193-209.

Recibido: 26 de junio de 2020

Aprobado: 1 de agosto de 2020

Conflictos de interés: No existen conflictos que declarar.

Contribución de autoría

1. Conceptualización: María Matilde García Lorenzo, Rafael Bello, Alejandro Rosete, Yaile Caballero
2. Curación de datos: Alejandro Rosete, María Matilde García Lorenzo,
3. Análisis formal: Alejandro Rosete, María Matilde García Lorenzo,
4. Adquisición de fondos: -
5. Investigación: María Matilde García Lorenzo, Yanela Rodríguez, Yaima Filiberto, Yaile Caballero, Rafael Bello
6. Metodología: María Matilde García Lorenzo, Rafael Bello
7. Administración del proyecto: -
8. Recursos: Yaima Filiberto
9. Software: Alejandro Ramón-Hernández, Beatriz Bello-García
10. Supervisión: -
11. Validación: Yanela Rodríguez, Alejandro Ramón-Hernández, Beatriz Bello-García
12. Visualización: -
13. Redacción – borrador original: María Matilde García Lorenzo, Alejandro Rosete, Yanela Rodríguez, Yaile Caballero, Rafael Bello
14. Redacción – revisión y edición: María Matilde García Lorenzo, Alejandro Rosete, Yanela Rodríguez, Yaile Caballero, Rafael Bello

Financiación: No hay fuentes de financiamiento que declarar además del salario devengado por los autores.

