



IDEAS EN CIENCIAS DE LA INGENIERÍA

Facultad de Ingeniería, Universidad Autónoma del Estado de México

Localización de policlínico para mejorar la cobertura de salud del Reparto Piedra Blanca

Interpretando Modelos de IA en Cáncer de Mama con SHAP y LIME

Diseño de rutas para estaciones de cargas eléctricas de vehículos eléctricos con algoritmos Metaheurísticos

Evaluación Sísmica del Tren Interurbano México-Toluca con simulaciones de carga viva y Ciencia de Datos

Un modelo de Red Bayesiana para datos cualitativos de Vaginosis Bacteriana en mujeres embarazadas





Universidad Autónoma del Estado de México

Consejo Editorial

Universidad Autónoma del Estado de México

Directora: Dra. Rosa María Valdovinos Rosas

Editora técnica: Dra. Mireya Salgado Gallegos

Diseñador: Dr. Javier Salas García

Universitat Jaume I

Editora científica: Dra. Angélica Guzmán Ponce

Comité Científico

Dra. Lourdes Loza Hernández - Universidad Autónoma del Estado de México

Dr. Francisco Héctor Bañuelos García - Universidad Autónoma del Estado de México

Dra. Juana Canul Reich - Universidad Juárez Autónoma de Tabasco

Dr. Ginés García Mateos - Universidad de Murcia, España

Dr. Álvaro Eduardo Lentz Herrera - Universidad Autónoma de la Ciudad de México

Dra. Edith Rosalba Salcedo Sanchez - Universidad Autónoma de Guerrero

Dr. Juan Manuel Esquivel Martínez - Universidad Autónoma de Guerrero

Dr. Manuel González de la Rosa - Tecnológico Nacional Campus Toluca

Dr. Agustín Bustos Rosales - Tecnológico de Estudios Superiores de Monterrey

Editor Invitado

Dr. Héctor Rodrigo Amezcua Rivera

Universidad Nacional Autónoma de México

IDEAS EN CIENCIAS DE LA INGENIERÍA, Año 4 , Vol. 2, Núm. 2, julio-diciembre 2024, es una publicación semestral editada por la Universidad Autónoma del Estado de México, Instituto Literario 100 Ote., Colonia Centro, Toluca, Estado de México, C.P. 50000, Tels. 7222-14-08-55, <https://hemeroteca.uaemex.mx/index.php/ideasingeneria>, ideasencienciai_fi@uaemex.mx. Directora: Rosa María Valdovinos Rosas, Facultad de Ingeniería, Ciudad Universitaria, Cerro de Coatepec s/n, Toluca, Estado de México, C.P. 50130, Tel. 7222140855. Reserva de Derechos al Uso Exclusivo no. 04-2023-033108322300-102, ISSN: 2992-7447, ambos otorgados por el Instituto Nacional del Derecho de Autor (INDAUTOR). Responsables de la última actualización de este número, Mireya Salgado Gallegos y Javier Salas García, Facultad de Ingeniería. Diseño de portada y Diseño Editorial, Javier Salas García. Fecha de última modificación: 12 de julio de 2024.

Las opiniones expresadas por los autores no necesariamente reflejan la postura del editor de la publicación. Se autoriza la reproducción total o parcial del contenido aquí publicado sin fines de lucro, siempre y cuando no se modifique, se cite la fuente completa y su dirección electrónica en términos de licencia aplicable. Esta obra está bajo una licencia de Creative Commons Reconocimiento-NoComercial-SinObraDerivada 4.0 Internacional (CC BY-NC).

Hecho en México, Universidad Autónoma del Estado de México (UAEMEX), 2024.



CONTENIDO

	Pág.
Carta del Editor Dr. Héctor Rodrigo Amezcua Rivera	2
Localización de policlínico para mejorar la cobertura de salud del Reparto Piedra Blanca Guillén Sánchez, M. J.; Loza-Hernández, L.	4
Interpretando Modelos de IA en Cáncer de Mama con SHAP y LIME Guzmán-Ponce, A.; López-Bautista, J.; Fernandez-Beltrán, R.	15
Diseño de rutas para estaciones de cargas eléctricas de vehículos eléctricos con algoritmos Metaheurísticos Ortiz-Aguilar, L.; Palacios-Ortega, M.; Xoca Orozco, L. A.	30
Evaluación Sísmica del Tren Interurbano México-Toluca con simulaciones de carga viva y Ciencia de Datos Valdovinos-Rosas, R.M.; Arellano-Magdaleno, O. D.; López-Díaz, J. L.; Delgado-Hernández, D. J.	46
Un modelo de Red Bayesiana para datos cualitativos de Vaginosis Bacteriana en mujeres embarazadas Bautista Hernández, C. M.; Canul Reich, J.; López Ramírez, C.; De la Cruz Hernández, E.	62

Carta del Editor

Dr. Héctor Rodrigo Amezcua Rivera



En este número, la revista “Ideas en Ciencias de la Ingeniería” presenta una serie de trabajos de investigación innovadores en diversas ramas de la ingeniería. Estos estudios se alinean estrechamente con las tendencias actuales en investigación, generando sinergias que impulsan el avance del conocimiento en ingeniería. Los artículos de este número exploran un espectro diverso de temas que van desde problemas de optimización hasta aplicaciones médicas de vanguardia con algoritmos de Inteligencia Artificial (IA), así como la seguridad estructural de la infraestructura. Todos estos temas

son abordados mediante enfoques que involucran, en distintos niveles y escalas, a la ciencia de datos como eje transversal para la solución de problemas.

El primer trabajo de este número es un claro ejemplo de cómo los desarrollos científicos y tecnológicos pueden beneficiar directamente a la sociedad. Los autores de este artículo abordan un problema clásico en la investigación de operaciones y la teoría de optimización: el problema de la localización de una sola instalación. En su estudio, determinan la ubicación óptima de un policlínico que será construido en Piedra Blanca, Cuba.

Es evidente que una de las tendencias más destacadas en la investigación en cualquier área de la ingeniería es el uso, desarrollo y aplicación de algoritmos basados en IA. Estos algoritmos permiten, entre otras cosas, descubrir relaciones entre fenómenos que, debido al gran volumen de información, resultaban muy difíciles de identificar con enfoques tradicionales.

Sin embargo, como pasa en muchos desarrollos tecnológicos, es conveniente comprender las limitaciones de cada algoritmo, como los posibles sesgos que pueden surgir debido a la naturaleza de las bases de datos utilizadas. Para mitigar este y otros problemas, la Inteligencia Artificial Explicable (XAI) se centra en técnicas que hacen que las predicciones de los algoritmos de IA sean transparentes y comprensibles para los seres humanos. Lo cual, sin duda, nos ayuda a comprender mejor el problema y a delimitar la aplicabilidad de cada algoritmo de IA. Siguiendo esta tendencia, el segundo artículo de este número presenta la aplicación de dos técnicas comunes de la XAI para interpretar un modelo de IA empleado para el diagnóstico de cáncer de mama tomando como datos de entrada imágenes histológicas del carcinoma ductal invasivo.

Otro problema de optimización con un beneficio social directo se aborda en el tercer artículo. En este trabajo, los autores utilizan un algoritmo metaheurístico para optimizar la

ubicación y distribución de estaciones de carga de vehículos eléctricos en el estado de Guanajuato con el enfoque de maximizar la cobertura de servicio.

El desarrollo de infraestructura de un país debe considerar, en su planificación, consideraciones sobre la seguridad estructural. especialmente en regiones con actividad sísmica significativa, como algunas partes de México. Por esta razón, el tema abordado por el cuarto trabajo es de alta relevancia, ya que los autores realizan un estudio para analizar la respuesta estructural del Tren Interurbano México-Toluca frente a un escenario sísmico específico, considerando las características tanto del sismo como del funcionamiento del tren.

Para concluir este número, se presenta otro artículo que explora una aplicación de un algoritmo de IA en el campo de la medicina. En este trabajo, los autores discuten el uso de una red bayesiana causal para la identificar que bacterias coexisten y cómo se relacionan en el contexto de un diagnóstico positivo de vaginosis bacteriana.

Se les invita a dar lectura a cada una de estas valiosas e interesantes contribuciones que abordan problemas reales de la sociedad a través de la sinergia entre la optimización, la inteligencia artificial y la ciencia de datos. En este número encontrarán ideas inspiradoras y aplicaciones prácticas que contribuirán significativamente a sus campos de estudio y práctica profesional.

Atentamente,

Dr. Héctor Rodrigo Amezcua Rivera
Editor Invitado
"Ideas en Ciencias de la Ingeniería"

Localización de policlínico para mejorar la cobertura de salud del Reparto Piedra Blanca

Localization of polyclinic to improve healthcare coverage in Piedra Blanca District, Cuba

Guillén Sánchez, M. J.¹,  Loza-Hernández, L.^{1*}  (*llozahernandez@yahoo.com)

Recibido: 16 de abril de 2024 Aceptado: 24 de mayo de 2024

RESUMEN

El Sistema Nacional de Salud de Cuba (SNS) cuenta con diversas instituciones cuya responsabilidad es asegurar el acceso equitativo y sin restricciones a todos los programas y servicios de salud, así como proporcionar una cobertura del 100 % para toda la población. El nivel primario de atención médica se dedica a abordar aproximadamente el 80 % de los problemas de salud de la población, ofreciendo acciones de promoción y protección de la salud. El Policlínico desempeña un papel central en este sistema de salud. En la provincia de Holguín, la dirección de salud trabaja en fortalecer los Policlínicos de la Familia, los cuales buscan acercar los servicios de salud a los diferentes núcleos poblacionales. El objetivo del trabajo es ubicar geográficamente

un policlínico en el reparto Piedra Blanca, en la ciudad cubana de Holguín. En este trabajo se aplica el método del Problema de Localización de una Sola Instalación (SFLP, por sus siglas en inglés Single Facility Location Problem) para determinar la localización del Policlínico, que será construido en el reparto Piedra Blanca de la ciudad de Holguín. El SFLP tiene como objetivo principal la minimización de la distancia total recorrida a cada punto de demanda, al cual se le asocia un peso relativo entre este y la distancia total existente al punto de localización de la nueva instalación.

Palabras clave: Localización de instalaciones, Problema de localización de una sola instalación, Sistema de salud

¹Facultad de Ingeniería, Universidad Autónoma del Estado de México. Paseo Universidad S/N. Toluca, Estado de México, México. CP. 50130.



ABSTRACT

The Cuban National Health System (SNS) has institutions that have the obligation to guarantee free and equal access to all health programs and services to 100 % of the population. The Cuban SNS guarantees access to comprehensive health services that are provided in outpatient and hospital units. The primary health level should provide solutions to approximately 80 % of the population's health problems and offer actions to promote and protect health. The Polyclinic is the main institution of the National Health System. The Holguín Provincial Health Directorate is working on the consolidation of the Family Polyclinics, which promote an approach of services to population centers. The objective of the work is geographically located a polyclinic in the Piedra Blanca neighborhood, in the Cuban city of Holguín. For the development of this work, the Single Facility Location Problem method is applied to determine the location of the Polyclinic, which will be built in the Piedra Blanca neighborhood of the city of Holguín. The main objective of the Single Facility Location Problem is to minimize the total distance traveled, in which a relative weight is associated with each demand point and the total distance between the demand node and the location point of the new facility.

Keywords: *Facility location, Single Facility Location Problem, Healthcare system*

INTRODUCCIÓN

El Sistema Nacional de Salud (SNS) de Cuba está compuesto por varias instituciones encargadas de asegurar el acceso equitativo y libre a todos los programas y servicios de salud, brindando una cobertura del 100 % a la población. Este acceso no se ve influenciado por el nivel de ingresos ni el origen social. El SNS cubano garantiza el acceso a servicios de salud integrales tanto en unidades ambulatorias como hospitalarias. El nivel primario tiene como objetivo abordar aproximadamente el 80 % de los problemas de salud de la población, proporcionando acciones de promoción y protección de la salud. Aunque estas actividades pueden llevarse a cabo en todas las unidades del SNS, se centran en los policlínicos y consultorios del médico de familia, que corresponden en su mayoría a unidades de subordinación municipal. En el nivel secundario, se atiende cerca del 15 % de los problemas de salud, centrándose en tratar a individuos que ya presentan cuadros clínicos para prevenir complicaciones y rehabilitar a los pacientes inmediatamente. La unidad básica en este nivel es el hospital de subordinación provincial. En el nivel terciario, se atiende aproximadamente el 5 % de los casos, y se ocupa de atender secuelas o complicaciones de enfermedades específicas. La atención se ofrece principalmente en hospitales especializados e institutos de subordinación nacional [1].

El objetivo del trabajo es ubicar geográficamente un policlínico en el reparto Piedra Blanca, en la ciudad cubana de Holguín, Cuba a través del uso de métodos cuantitativos, que permitan asegurar que los

pacientes recorren la mínima distancia para recibir el servicio de salud requerido. La Figura 1 muestra la ubicación geográfica del Reparto donde es necesario localizar el nuevo policlínico.

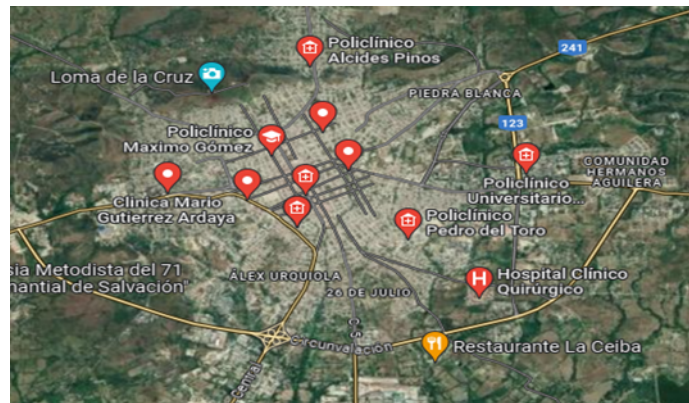


FIGURA 1. *Policlínicos familiar, ciudad de Holguín, Cuba. Fuente: Elaboración propia.*

En la literatura existen varios métodos cuantitativos para determinar la mejor ubicación de instalaciones con base en los requerimientos de los tomadores de decisiones y necesidades propias de las empresas de los diferentes sectores; para el caso del sector salud, estos métodos se usan como herramientas para determinar la mejor ubicación de instalaciones médicas (hospitales, centros de salud, clínicas, etc.). Para determinar la mejor ubicación de estas instalaciones, son considerados criterios que identifican las mejores localizaciones, para el caso de estudio, el principal criterio es la necesidad de la población que atenderá dicha instalación, por lo tanto, herramientas como Centro P, Cobertura Máxima, Problema de Localización de una Sola Instalación y Centro de Gravedad son algunos de los métodos que se pueden utilizar para lograr el objetivo planteado. Para el desarrollo de este trabajo, se aplica el método del Problema de Localización de una Sola Instalación (SFLP) para determinar la localización del Policlínico, que será construido en el reparto Piedra Blanca de la ciudad de Holguín, Cuba. Específicamente se seleccionó el método (SFLP) debido a que éste permite considerar las características geográficas de las áreas en estudio; cabe mencionar que la Zona de Estudio 1 se encuentra en una zona montañosa, lo que resulta en una densidad poblacional menor en comparación con otras zonas analizadas. Como parte de las características geográficas, a la densidad poblacional se le asigna un peso relativo asociado a cada punto de demanda o zona de estudio.

PROBLEMA DE LOCALIZACIÓN PARA UNA SOLA INSTALACIÓN

Actualmente, el Policlínico se erige como la institución fundamental en el marco del SNS a nivel nacional, el cual dirige las instituciones de salud en su área respectiva. Su responsabilidad principal es la integración y ejecución de todas las acciones del sistema de salud, siendo encargado de guiar al paciente a lo largo de todos los niveles de atención hasta la resolución de sus problemas de salud. La organización

y planificación de los servicios del centro de salud se llevan a cabo considerando los problemas de salud de la población o del área territorial que atiende. Además, se asegura un programa continuo de capacitación y formación para los recursos humanos, con el fin de garantizar el funcionamiento eficiente de los servicios y abordar otras necesidades, en colaboración con las sedes universitarias. De manera sistemática, se evalúa la satisfacción tanto de los proveedores de servicios como de los usuarios, mediante el departamento de atención a la población y se realiza una exploración constante desde los diferentes servicios utilizando tantos métodos regulares como de búsqueda activa, con el objetivo de tomar decisiones informadas en la organización de los servicios.

El Problema de Localización de una Sola Instalación fue estudiado por primera vez por Fermat como una de las ramas del problema de ubicación [2]. Esta rama del problema de la localización de instalaciones ha atraído el interés de muchos investigadores desde entonces, muestra de ello, en Irán en 2022 se realizó un estudio sobre el problema inverso de ubicación de una sola instalación, equilibrando la distancia del proveedor a los clientes [3]. El problema de ubicación bajo incertidumbre de la demanda fue propuesto por Jaehyeon y Park [4]. Gökhan aborda el problema de la ubicación de una sola instalación en múltiples regiones con diferentes normas y luego idearon un algoritmo de ramificación y unión especialmente diseñado para resolver el problema [5], solo por mencionar algunos.

El problema de la ubicación de una sola instalación es de los problemas científicos de optimización y gestión más importantes. Es utilizado para modelar problemas del mundo real, como determinar la ubicación de oficinas administrativas, militares, centros educativos, estaciones de bomberos, sucursales bancarias, hospitales, etc., lo cual justifica el amplio interés de expertos en el área, presentando resultados valiosos [6]. Sin embargo, la existencia de ambigüedad e incertidumbre en el mundo real es un hecho innegable, por lo tanto, en muchos casos, es imposible determinar los valores exactos de los parámetros; en tales casos, se utiliza la teoría difusa, que es una herramienta adecuada para expresar y describir la incertidumbre y la precisión de los eventos [7]. El problema de la ubicación difusa también ha recibido una atención significativa por parte de muchos investigadores y numerosos estudios se han realizado en esta área. En 2013, se presentó un estado del arte sobre la metodología de simulación para problemas de diseño de instalaciones [8]. Jiuping y Song estudiaron el problema de diseño dinámico multiobjetivo para instalaciones de construcción temporales con departamentos de áreas desiguales en un entorno difuso [9].

Analizando lo anteriormente expuesto y teniendo en cuenta las características y objetivos del sistema de salud cubano, se hace necesario acercar el sistema de salud a los habitantes de cada localidad. El modelo SFLP ha sido utilizado en trabajos anteriores para localizar instalaciones hospitalarias, ya que tiene como objetivo principal la minimización de la distancia total recorrida a la cual se asocia un peso relativo (W_i) a cada punto de demanda, donde i es el número de puntos de demanda (m) (este peso relativo se puede calcular en función de las distintas variables de interés) y la distancia total existente

entre el nodo de demanda (p_i) y el punto de localización de la nueva instalación x la cual se denota por ($d(x, p_i)$). La Ec. (1) muestra de forma matemática el planteamiento descrito [11]:

$$F.O.MIN F(X) = \sum_{i=1}^m W_i d(x, p_i) \quad \forall i = 1, 2, \dots, m \quad (1)$$

Donde:

$d(x, p_i)$ es la distancia que puede ser calculada ya sea rectilínea, cuadrada rectilínea o euclidiana desde el punto de demanda (p_i) hasta la zona de posible ubicación x .

W_i es el peso relativo asociado a cada punto de demanda i [10].

De acuerdo con esto, se propone el modelo de SFLP con el objetivo de ubicar la instalación de manera óptima, minimizando la distancia total recorrida [12].

Este método ha sido utilizado para ubicar instalaciones médicas como fue el caso de la localización geográfica de una instalación hospitalaria para mejorar la cobertura en salud de las localidades de Sumapaz, Usme y Rafael Uribe en Colombia en el año 2019.

CASO DE ESTUDIO

El objetivo de este trabajo es ubicar geográficamente un policlínico en el reparto Piedra Blanca, en la ciudad cubana de Holguín, Cuba a través de la aplicación del método SFLP, que permita asegurar que los pacientes recorran la mínima distancia para recibir el servicio de salud requerido. La Figura 1 muestra la localización de Piedra Blanca, la cual requiere la instalación del nuevo policlínico. Para analizar el caso en cuestión, se subdividió la zona en tres regiones (Figura 2), atendiendo a la densidad poblacional y la accesibilidad a cada zona.



FIGURA 2. Subdivisión de Zona de estudio. Fuente: Elaboración propia.

El país experimenta una transición demográfica, caracterizada por una fecundidad muy baja (1.5 hijos por mujer), una mortalidad reducida (7.3) y el aumento en la esperanza de vida al nacer, que alcanzó los 77 años en 2022. Estos fenómenos demográficos han resultado en una disminución en la tasa de crecimiento poblacional (de 4.6 en 1997 a -0.2 en 2007) y un envejecimiento de la población. Actualmente, más del 16 % de la población tiene 60 años o más, mientras que solo el 18.4 % tiene menos de 15 años. En la Tabla 1 se muestran los datos demográficos de la zona de estudio [1].

Tabla 1. Datos demográficos de la zona de estudio (año 2022). Fuente: Elaboración Propia.

Zona	Densidad Poblacional (hab/km ²)	Extensión (km ²)	Habitantes (hab)
1	678	0.380412	1782
2	1267	0.478788	2646
3	989	0.806752	1226

De acuerdo con el relieve, la disposición de las vías y la disposición del suelo se identificaron varias zonas en las cuales puede ubicarse la instalación hospitalaria. Se generó sobre el mapa del reparto Piedra Blanca una malla con divisiones de 100 metros por 100 metros, en las zonas identificadas como factibles, y se encontró el centroide de estas zonas, ver Tabla 2.

Tabla 2. Coordenadas de zonas candidatas *k*. Fuente: Elaboración propia.

Ubicación	Coordenada centroide X	Coordenada centroide Y
1	20.898452	-76.247565
2	20.892396	-76.24554
3	20.895692	-76.243143
4	20.897407	-76.242767
5	20.894302	-76.242127
6	20.898275	-76.242127
7	20.898806	-76.238728
8	20.898555	-76.238415

A través del Plan Director de la Ciudad y la malla realizada, se obtuvieron 8 zonas posibles, como se muestra en la Figura 3 (las zonas posibles están enmarcas en cuadrículas rojas).

Teniendo en cuenta la población de cada zona del Reparto (Tabla 1), la localización del centroide de cada zona de demanda se muestra en la Tabla 3, las ubicaciones posibles para la instalación del



FIGURA 3. *Zonas de Posible Ubicación. Fuente: Elaboración propia.*

policlínico (Tabla 2 y Figura 3) y las disposiciones del Plan Director de la ciudad en cuanto al uso de suelos, se procedió a definir los parámetros del modelo. El peso relativo o importancia de cada zona resulta de calcular la proporción entre la población de cada zona y la población total del reparto [13] .

Tabla 3. Coordenadas de las zonas de estudio *i*. Fuente: Elaboración propia.

Zona	Coordenada centroide X	Coordenada centroide Y
1	20.90235405	-76.24166502
2	20.8964464	-76.2427995
3	20.89756683	-76.24051708

Formulación del problema al caso de estudio

i: índice que identifica las zonas de demanda, donde $i=1,2,3$.

k: índice que identifica las zonas habilitadas para la localización del policlínico, donde $k=1,2,\dots,8$

X_i : abscisa del centroide de la zona de demanda *i*, donde $i=1,2,3$

Y_i : ordenada del centroide de la zona de demanda *i*, donde $i=1,2,3$

W_i : peso relativo de cada zona de demanda *i*, donde $i=1,2,3$

X_k : abscisa del centroide de la zona habilitada *k* para la ubicación del policlínico, donde $k=1,2,\dots,8$

Y_k : ordenada del centroide de la zona habilitada *k* para la ubicación del policlínico, donde $k=1,2,\dots,8$

Se definieron las variables de decisión del modelo, entre las cuales se incluyen las coordenadas del punto donde se localizará la instalación, así como las distancias rectilíneas recorridas desde el nuevo hospital hasta los nodos de demanda.

Variables de decisión

XF : Abscisa de la nueva instalación hospitalaria

YF : Ordenada de la nueva instalación hospitalaria

DX_i : Distancia rectilínea en el eje X desde los centroides de las zonas tipo i a la nueva instalación hospitalaria, donde $i=1,2,3$

DY_i : Distancia rectilínea en el eje Y desde los centroides de las zonas i a la nueva instalación hospitalaria, donde $i=1,2,3$

M_k : Variable binaria que indica si el hospital es instalado en la zona k , donde $k=1,2,\dots,8$

Una vez determinados los parámetros y las variables de decisión, se procede a la elaboración del modelo canónico.

Función Objetivo:

$$\text{MIN } F = \sum_{i=1}^3 W_i * (DX_i + DY_i) \quad (2)$$

Sujeto a:

$$DX_i + XF = X_i \quad \forall i \quad (3)$$

$$DY_i + YF = Y_i \quad \forall i \quad (4)$$

$$\sum_{k=1}^8 M_k = 1 \quad (5)$$

$$\sum_{k=1}^8 X_k * M_k = XF \quad (6)$$

$$\sum_{k=1}^8 Y_k * M_k = YF \quad (7)$$

$$M_k \in \{1, 0\} \quad \forall k \quad (8)$$

La función objetivo Ec. (2) busca minimizar la distancia total ponderada recorrida desde los nodos de demanda hasta la localización del nuevo policlínico; las restricciones Ec. (3) y Ec. (4) cuantifican la distancia rectilínea recorrida tanto en el eje de las abscisas como en el de las ordenadas desde cada nodo de demanda hasta el nuevo policlínico; por su parte, las restricciones Ec. (6) y Ec. (7) garantizan que la localización del nuevo policlínico sea en la zona candidata que se activa, siendo estas activadas gracias a la restricción Ec. (5), que garantiza que solo una de las zonas albergue el policlínico a instalar. Finalmente, la Ec. (8) valida el valor binario de la variable de decisión M_k .

RESULTADOS OBTENIDOS

Una vez diseñado el modelo se busca la solución al problema planteado. Para ello, se hizo uso del Software GAMS SIDE versión 23.5 y se codificaron tanto la función objetivo como las restricciones con la información relevante recolectada para la investigación, con el fin de obtener posibles alternativas de solución a la localización de la nueva instalación hospitalaria (ver Figura 4).

UBICACIÓN DE POLICLÍNICO			
Execution			
----	90 VARIABLE F.L	=	774273.28
	VARIABLE XF.L	=	20.897
	VARIABLE YF.L	=	76.242
	Distancia		
	COORDENADA X DEL PUNTO DONDE SE UBICA LA INSTALACION		
	COORDENADA Y DEL PUNTO DONDE SE UBICA LA INSTALACION		
----	90 VARIABLE M.L	VARIABLE BINARIA QUE INDICA LA ZONA k DONDE SE UBICA LA INSTALACION	
UBICACION1	4.000		
EXECUTION TIME	=	0.031 SECONDS	3 Mb WIN235-235 Jul 2, 2010

FIGURA 4. Resultados arrojados por programa GAMS SIDE. Fuente: Elaboración propia.

El modelo fue ejecutado en una computadora con procesador Intel(R) Celeron (R) CPU N3050 @160GHz con memoria RAM de 4GB. La solución tardó en obtenerse 0.031 segundos. Los resultados fueron F= 774273.286, la ubicación seleccionada fue la número 4, con distancias totales en X y en Y de 555.643 y 338.440 metros respectivamente, la ubicación del nuevo policlínico será X=20.897407 Y=-76.242767, lo que puede expresarse en coordenadas geográficas como: 20°53'50.7"N 76°14'34.0"W; ver Figura 5.

Los resultados obtenidos son viables, debido a que la localización en el noreste de la región delimitada para la solución del problema es un área montañosa que por su relieve presenta baja densidad poblacional



FIGURA 5. *Ubicación geográfica del policlínico. Fuente: Elaboración propia.*

y por tanto no sería una ubicación adecuada para el servicio de salud.

CONCLUSIONES

Es relevante resaltar que el objetivo general de la investigación se materializó efectivamente, el cual era determinar una propuesta de localización geográfica de un policlínico para solucionar los problemas de cobertura de salud en el Reparto Piedra Blanca, Holguín, Cuba; haciendo uso de métodos cuantitativos de localización de instalaciones, considerando la densidad poblacional de la zona, para así llegar a la mayor cantidad de población posible. Los beneficios inherentes al uso de esta herramienta son reducir las distancias a recorrerse entre los nodos de demanda y el centro médico y que se analiza, al mismo tiempo, el impacto de otras variables como en este caso lo es la densidad poblacional; logrando los objetivos planteados del desarrollo del trabajo.

REFERENCIAS

- [1] E. Domínguez-Alonso, M. P. Eduardo Zacea. Sistema de salud de Cuba. 2009.
- [2] R. F. Love, J. James, G. Morris, and G. O. Wesolowsky. Facilities Location: Models & Methods. North-Holland Publishing Co., New York, 1988.
- [3] S. Omid, J. Fathali. Inverse single facility location problem on a tree with balancing on the distance. *Journal of Industrial and Management Optimization*, 12(5), 2022. DOI: 10.3934/jimo.2021017
- [4] J. Ryu, S. Park. A branch-and-price algorithm for the robust single-source capacitated facility location problem under demand uncertainty. *EURO Journal on Transportation and Logistics*, 11(26):100069, 2022. DOI: 10.1016/j.ejtl.2021.100069

- [5] G. A. Altay, M. H. Akyüz, T. Öncan. Solving a minisum single facility location problem in three regions with different norms. *Annals of Operations Research*, 321(1), 2023. DOI: 10.1007/s10479-022-04952-5
- [6] M. J. Varnamkhasti. Overview of the algorithms for solving the P-median facility location problems. *Advanced Studies in Biology*, 4(2):49-55, 2012.
- [7] N. A. Taghi-Nezhad. Fuzzy linear fractional programming for container transportation optimization. *International Journal of Industrial Mathematics*, 5(4):309-316, 2023.
- [8] C. N. Kulkarni, M. I. Talib, R. S. Jahagirdar. Simulation methodology for facility layout problems. *The International Journal of Engineering and Science*, 2(2):24-30, 2013.
- [9] J. Xu, X. Song. Multi-objective dynamic layout problem for temporary construction facilities with unequal-area departments under fuzzy random environment. *Knowledge-Based Systems*, 81(2), 2015. DOI: 0.1016/j.knosys.2015.02.001
- [10] R. Zanjirani Farahani. Facility Location: Concepts, Models, Algorithms and Case Studies. Heidelberg: Physica-Verlag, 2009. DOI: 10.1007/978-3-7908-2151-2
- [11] R. Zanjirani Farahani, Z. Drezner, N. Asgari. Single facility location and relocation problem with time dependent weights and discrete planning horizon. *Annals of Operations Research*, 167:353-368, 2009.
- [12] A. F. Vélez Osorio, C. V. García Prieto. Localización geográfica de una instalación hospitalaria para mejorar la cobertura en salud de las localidades de Sumapaz, Usme y Rafael Uribe Uribe. 2019. Recuperado el 10 de junio de 2023, de https://ciencia.lasalle.edu.co/ing_industrial/136
- [13] Ministerio de Salud Pública de Cuba. Anuario Estadístico. La Habana: MINSAP, 2022.

Interpretando Modelos de IA en Cáncer de Mama con SHAP y LIME

Interpreting AI Models in Breast Cancer with SHAP and LIME

Guzmán-Ponce, A.^{1*}, López-Bautista, J.², Fernandez-Beltrán, R.³ (*aguzman@uji.es)

Recibido: 31 de mayo de 2024 Aceptado: 17 de junio de 2024

RESUMEN

La Inteligencia Artificial (IA) ha transformado múltiples disciplinas, en concreto en el área médica se ha distinguido por su capacidad de mejorar diagnósticos y tratamientos. No obstante, la falta de transparencia en los sistemas ha generado desafíos prácticos. La Inteligencia Artificial Explicable (XAI) ha surgido como una solución para abordar estos problemas, permitiendo a los profesionales de la salud comprender y verificar las decisiones de los algoritmos. En este trabajo se aplicarán dos técnicas comúnmente usadas de XAI, específicamente SHapley Additive

exPlanations (SHAP) y Local Interpretable Model-Agnostic Explanations (LIME), para interpretar un modelo de red neuronal convolucional usado en el diagnóstico de cáncer de mama a partir de imágenes histológicas del Carcinoma Ductal Invasivo (IDC). Los resultados muestran que ambas técnicas son efectivas para identificar las contribuciones individuales de las características en las predicciones del modelo, resaltando que LIME ofrece una granularidad detallada a nivel de píxel.

Palabras clave: *Inteligencia Artificial Explicable, SHAP, LIME, Carcinoma Ductal Invasivo*

ABSTRACT

Artificial Intelligence (AI) has transformed multiple disciplines, particularly medicine, where it has been distinguished by its ability to improve

diagnostics and treatments. However, the lack of transparency in the systems has raised practical challenges. Explainable Artificial Intelligence

¹Departamento de Lenguajes y Sistemas Informáticos, Instituto de Nuevas Tecnologías de Imagen, Universitat Jaume I, 12071 Castelló de la Plana, España.

²Hospital Regional ISSEMyM Toluca, Médico Adscrito al Departamento de Anatomía Patológica, Calle Dr. Nicolás San Juan S/N, Delegación San Lorenzo Tepaltitlán I, Delegación San Lorenzo Tepaltitlán, 50010 Santa Cruz Atzacapotzaltongo, Méx.

³Departamento de Informática y Sistemas, Universidad de Murcia, Murcia, España.



(XAI) has emerged as a solution to address these issues, allowing healthcare professionals to understand and verify the decisions made by algorithms. In this work, two commonly used XAI techniques, Shapley Additive explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), are applied to interpret a convolutional neural network model used in the diagnosis of breast cancer from histological images of Invasive Ductal Carcinoma (IDC). The results show that both techniques effectively identify the individual contributions of features in the predictions of the model, highlighting that LIME offers detailed granularity at the pixel level.

Keywords: *Explainable Artificial Intelligence, SHAP, LIME, Invasive Ductal Carcinoma*

INTRODUCCIÓN

En las últimas décadas, la Inteligencia Artificial (IA) ha revolucionado disciplinas desde la agricultura hasta la biotecnología, destacando su impacto en medicina por mejorar diagnósticos y tratamientos. La IA ha sido clave en el reconocimiento de patrones y la predicción de comportamientos, como se evidencia en su aplicación durante la pandemia de COVID-19, donde facilitó la toma de decisiones informadas al prever dinámicas virales [1]. Además, en radiología, las técnicas de IA han permitido distinguir eficazmente entre imágenes de rayos X de pacientes con COVID-19 y otras enfermedades respiratorias, demostrando su eficacia [2] [3].

La implementación de la IA en medicina ha facilitado avances notables en análisis y diagnóstico, mejorando la precisión y eficacia de los tratamientos [4]. No obstante, la falta de transparencia en los procesos de decisión de muchos sistemas de IA presenta retos tanto éticos como prácticos. En el área médica es importante contar con un entendimiento claro de cómo los algoritmos toman decisiones, dada la necesidad de justificar cada diagnóstico o recomendación de tratamiento de manera detallada y comprensible [5].

Ante esta necesidad, ha surgido la Inteligencia Artificial Explicable (eXplainable Artificial Intelligence, XAI, por sus siglas en inglés), como una respuesta a la “caja negra” que se genera en los modelos de IA [6]. XAI permite a los profesionales de la salud comprender y verificar los procesos mediante los cuales se llegan a tomar decisiones automáticas, fomentando así la confianza y facilitando la integración de estos sistemas en las prácticas médicas regulares [7].

A pesar del esfuerzo que la comunidad científica ha realizado, existe una falta de estudios en la aplicación de técnicas de XAI específicamente en imágenes de histología del cáncer de mama IDC. Derivado de este hecho, este trabajo explora cómo diferentes metodologías, como SHAP y LIME, pueden optimizar la interpretación y el análisis de imágenes de histología. La pregunta de investigación central de nuestro estudio se centra en determinar cuál de estas técnicas de XAI proporciona mayor claridad en

el diagnóstico del cáncer de mama.

INTELIGENCIA ARTIFICIAL EXPLICABLE

XAI se ha utilizado para hacer que los procesos de toma de decisiones de la IA sean transparentes y comprensibles, especialmente en campos críticos como la salud, donde las metodologías de IA a menudo resultan complejas para muchos profesionales médicos. XAI busca responder a la pregunta: ¿Qué hace que esta predicción sea confiable? [8]. Por lo tanto, se considera una técnica analítica posterior al proceso de extracción de conocimiento en las metodologías convencionales de IA. Su objetivo es destacar las características que influyen en el rendimiento del modelo, aumentando así la confianza en los sistemas de IA y facilitando su integración en áreas sensibles como la medicina [4].

Existen dos técnicas de XAI agnósticas al modelo de IA, lo que significa que pueden aplicarse a cualquier tipo de modelo de extracción de conocimiento, sin importar su estructura interna. Las técnicas de Local Interpretable Model Agnostic Explanations (LIME) y SHapley Additive exPlanations (SHAP) son métodos ampliamente utilizados [6]. Ambos enfoques se centran en proporcionar interpretaciones locales de las predicciones, ayudando a entender cómo cada característica contribuye a una decisión específica del modelo.

SHapley Additive exPlanations

SHAP se basa en los valores de Shapley, que cuantifica la contribución de cada característica a la predicción de un modelo. Si se considera un conjunto de datos DS con n instancias, cada una representada por un par (x, y) , donde x es un arreglo de m características $[x_1, x_2, \dots, x_m]$, y es la clase. Dado M el conjunto de todas las características. El valor de Shapley para la i -ésima característica se define como la Ec. (1):

$$\phi_i = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|! (|M| - |S| - 1)!}{|M|!} [f(S \cup \{i\}) - f(S)] \quad (1)$$

Donde S es un subconjunto de M sin la característica i -ésima; $f(S \cup \{i\})$ es la predicción del modelo con la característica i y $f(S)$ es la predicción sin ella [8].

Para aplicar la fórmula, primero se calculan todas las posibles combinaciones de características excluyendo la i -ésima característica. Luego, se evalúan las predicciones del modelo con y sin la característica i para cada combinación. La diferencia en las predicciones se multiplica por un peso que depende del tamaño del subconjunto y del total de características. Finalmente, se suman todas estas contribuciones ponderadas para obtener el valor SHAP, que representa la contribución promedio de la característica i a lo largo de todas las posibles combinaciones. Las características que aumentan las predicciones generan

valores SHAP positivos, mientras que las que disminuyen las predicciones tienen valores negativos.

Local Interpretable Model Agnostic Explanations

LIME se basa en la aproximación local del modelo en la vecindad de una instancia específica. La explicación de LIME para la instancia x se define en la Ec. (2):

$$\xi(x) = \arg \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (2)$$

Donde G son las posibles explicaciones; f es el modelo original que se quiere explicar; g es el modelo local explicable; L es la pérdida que mide la cercanía de la explicación a la predicción del modelo original; π_x es una medida de proximidad que pondera las instancias de acuerdo con su similitud con x . $\Omega(g)$ es una medida de la complejidad del modelo de explicación $g \in G$, por ejemplo, el número de características de entrada.

Para aplicar LIME, primero se generan muestras locales perturbando de manera controlada la instancia x original. Luego, se evalúa el modelo en estas muestras perturbadas para obtener las predicciones correspondientes. A continuación, se asignan pesos a estas muestras usando una medida de proximidad. Con las muestras ponderadas, se entrena un modelo explicable g . Finalmente, se optimiza este modelo para minimizar la pérdida ponderada y la penalización de complejidad, obteniendo así una aproximación localmente interpretable del modelo completo. La interpretación de los valores de LIME es similar a la de SHAP. Las características que aumentan las predicciones generan valores LIME positivos, mientras que las que disminuyen las predicciones tienen valores LIME negativos.

XAI EN IMÁGENES MÉDICAS

En la investigación médica, especialmente en el diagnóstico por imágenes Borys et al. [9], resaltan cómo la disponibilidad de datos y los avances en el alto rendimiento computacional han aumentado el potencial de la IA, pero también señalan la necesidad de entender sus procesos internos en contextos críticos. El estudio se enfoca en métodos de XAI basados en la saliencia, es decir, la cualidad por la cual ciertas partes de una imagen atraen más la atención que otras, son cruciales en la imagenología médica, ya que permiten proporcionar explicaciones directamente sobre los datos de entrada (imágenes), mejorando la confianza y la comprensión de los resultados de IA por parte de médicos y pacientes.

Por otro lado, Shovon et al. [10] abordan la clasificación de imágenes histopatológicas de cáncer de mama HER2, un subtipo de cáncer que se caracteriza por la sobreexpresión del receptor del factor de crecimiento epidérmico humano 2, usando un modelo que integra una red DenseNet201 que optimiza el entrenamiento al conectar densamente cada capa con todas las demás, reduciendo el sobreajuste; y

Xception que utiliza convoluciones separables en profundidad para procesar eficientemente las características. Para interpretar los resultados implementaron Grad-CAM el cual usa gradientes de la última capa convolucional para crear mapas de calor que destacan regiones críticas en imágenes.

Mientras que, Kaplun et al. [11] desarrollaron un método basado en los momentos de Zernike (son polinomios ortogonales en el círculo unitario, útiles para análisis de imágenes por su invarianza rotacional) para caracterizar las imágenes de células cancerosas de mama y posteriormente usar una red neuronal simple para la clasificación. Integraron un modelo local de interpretación, lo que permitió que los resultados obtenidos por el modelo de clasificación de imágenes de células cancerosas fueran más sencillos de interpretar.

METODOLOGÍA

Las imágenes histológicas son imágenes microscópicas de tejidos biológicos que han sido especialmente preparados y teñidos para revelar detalles estructurales, siendo fundamentales en el campo de la patología, ya que permiten examinar la arquitectura microscópica de los tejidos para diagnosticar enfermedades, evaluar la severidad de las condiciones patológicas y entender la patogénesis de diversas enfermedades. En la actualidad, este tipo de imágenes han sido usadas en arquitecturas de aprendizaje profundo donde se pueden aplicar tareas de detección, segmentación y clasificación de tejidos con el objetivo de producir resultados comparables o superiores a los enfoques basados en características diseñadas manualmente, tradicionales en la patología [12].

Conjunto de datos

Para el análisis del objetivo de este artículo, se emplea un conjunto de datos compuesto por imágenes histológicas, centrado específicamente en el estudio del Carcinoma Ductal Invasivo (IDC) sin patrón específico NOS (Not Otherwise Specified), la variante más común de cáncer de mama, que representa aproximadamente el 70-80 % de todos los casos de esta enfermedad [12]. Este tipo de cáncer se distingue por la invasión de las células neoplásicas hacia el estroma mamario causando una reacción llamada desmoplasia.

El conjunto tomado de [13] está compuesto de imágenes en fragmentos de 50x50 píxeles teñidos con Hematoxilina y Eosina (H&E), que es una técnica de tinción común utilizada en histopatología para destacar las características de las células y tejidos. Cada imagen se encuentra etiquetada como clase 1 si contienen células IDC (IDC +), en caso contrario se etiqueta como clase 0 (IDC -), esta etiqueta fue asignada manualmente [12]. Teniendo un total de 5547 imágenes, que se distribuyen en 2759 imágenes con ausencia de IDC y 2788 con presencia de IDC. La Figura 1 muestra un ejemplo representativo para cada categoría.

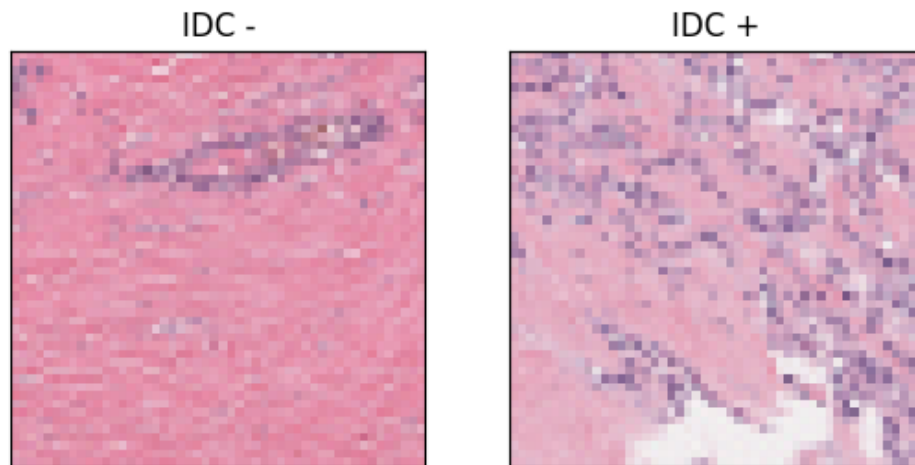


FIGURA 1. *Imágenes de tejido mamario H&E, a la izquierda con ausencia de cáncer, a la derecha con presencia de cáncer.*

Modelo de Aprendizaje

Para la extracción de conocimiento, se implementó un modelo de red neuronal convolucional (CNN) usado en [12], [13], como se observa en la Figura 2. El modelo se compone de diversas capas que ayudan a capturar características esenciales de las imágenes de entrada para realizar la clasificación de IDC. Inicialmente, se utilizan dos capas de convolucionales con 32 y 64 filtros respectivamente, seguidas cada una por una función de activación ReLU que introduce no linealidades al modelo. Luego, una capa de agrupación reduce la dimensionalidad de los datos, lo que ayuda a hacer el modelo más eficiente y a reducir el sobreajuste.

Para combatir aún más el sobreajuste, se incluyen capas de Dropout que, aleatoriamente, ignoran algunos nodos durante el entrenamiento, lo que ayuda a mejorar la generalización del modelo en datos no vistos. Las últimas capas convierten los mapas de características multidimensionales en un vector unidimensional, permitiendo que las capas densas procesen estas características para realizar la clasificación final, el uso de la función de activación softmax distribuye las probabilidades a través de las clases definidas.

El modelo emplea el optimizador Adadelta, adecuado para datos con características dinámicas por su adaptabilidad en el tamaño de pasos de aprendizaje. Se configuró un lote de 128 y se entrenó durante 200 épocas, las capas de Dropout tienen un porcentaje de 0.25 después de la segunda capa convolucional y 0.5 después de la capa densa.

El conjunto de datos fue dividido en 80 % para entrenamiento y el 20 % restante para la fase de prueba.

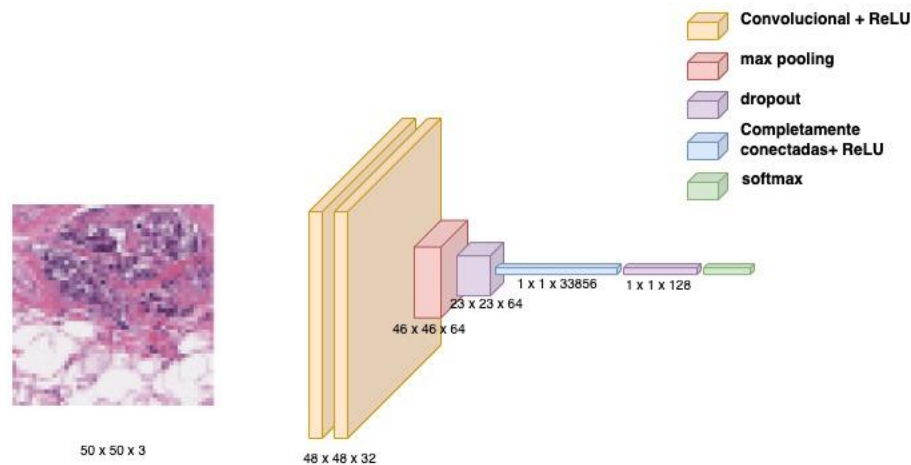


FIGURA 2. Diagrama de la red neuronal convolucional para la clasificación de imágenes histológicas.

La evaluación del modelo se realizó utilizando como métrica la exactitud (accuracy) definida como la proporción de predicciones correctas (tanto verdaderos positivos como verdaderos negativos) respecto al total de casos evaluados.

XAI

Una vez que el modelo CNN ha sido entrenado, es posible aplicar métodos que permitan identificar las porciones que han sido más influyentes en la decisión de clasificación de la red. Esta tarea se enmarca en el campo de XAI mediante SHAP y LIME, estas herramientas facilitan la interpretación de las decisiones del modelo, ayudando a identificar posibles sesgos en los datos o en el aprendizaje del modelo.

RESULTADOS

El modelo CNN empleado para la clasificación de imágenes de IDC alcanzó una precisión del 75 %, condicionada por una resolución de 50x50 píxeles y variaciones en la tinción H&E. Aunque la tinción H&E es estándar en histopatología, sus variaciones limitan significativamente la captura de características celulares en el modelo utilizado.

XAI en IDC

Una vez evaluado el modelo, se deduce que existen oportunidades para mejorar su desempeño mediante la optimización de los hiperparámetros, tales como el número y tamaño de los filtros en las capas convolucionales. En línea con el objetivo de este trabajo, los resultados del análisis de las percepciones sobre la contribución individual de cada característica en las predicciones del modelo es especialmente útil para:

1. Casos en los que el modelo ha clasificado incorrectamente una instancia, identificando qué características influyen negativamente en la decisión del modelo y cómo podrían ajustarse.
2. Casos en los que el modelo ha clasificado correctamente una instancia, entendiendo las características que contribuyen positivamente a las decisiones acertadas.

La metodología SHAP utiliza una escala cromática en la que el color azul simboliza una disminución en la probabilidad de la clase asociada, orientada específicamente hacia la ausencia de cáncer, con una intensidad que es directamente proporcional al impacto de la característica evaluada. En contraste, la metodología LIME emplea el color verde para indicar características que mejoran positivamente la precisión de las predicciones, mientras que el color rojo señala aquellos atributos que tienen un efecto negativo.

La Figura 3 muestra que hay áreas para SHAP con píxeles rojos que deberían incrementar la predicción hacia la presencia de cáncer, pero el modelo ha predicho incorrectamente la clase. Esto sugiere que, a pesar de la presencia de factores que indican cáncer, las zonas no coinciden con las infiltraciones del carcinoma, por lo que otros factores, como los píxeles azules, han llevado al modelo a una conclusión errónea, dado que coinciden en zonas con ausencia de infiltración. Esto puede deberse al sesgo del modelo, dada la representación insuficiente en los datos de entrenamiento de características similares.

La Figura 4 muestra la explicación generada por LIME para una imagen histológica incorrectamente clasificada como ausente de cáncer, cuando en realidad tiene presencia de IDC. Esta representación visual evidencia que el modelo ha asignado de manera uniforme el color verde a todos los píxeles de la imagen, lo que indica que ha interpretado estas características como señales de ausencia de cáncer, contribuyendo

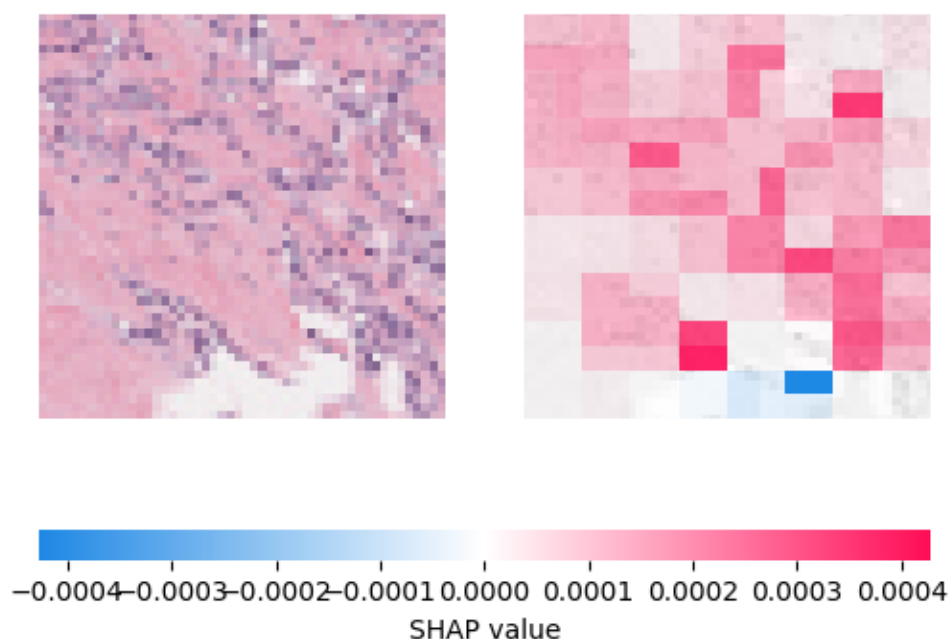


FIGURA 3. Valores SHAP para una imagen clasificada erróneamente como sin cáncer, cuando en realidad la etiqueta correcta indica la presencia de cáncer.

así erróneamente a la predicción de la clase. Esta observación sugiere la necesidad de reentrenar el modelo incorporando un volumen ampliado de imágenes, con el objetivo de mejorar su capacidad de discriminación y precisión en la clasificación de imágenes histológicas.

En la Figura 5 se observan los valores de SHAP correspondientes a una imagen correctamente clasificada con ausencia de cáncer. Los valores marcados en azul se localizan en áreas donde no hay reacción a la tincura, indicando la ausencia de cáncer.

De la Figura 6 se observa que los píxeles marcados en verde, correspondientes a LIME, indican zonas donde no se detecta la infiltración del carcinoma.

Un caso particular se ilustra en la Figura 7, donde el modelo predijo la presencia de cáncer, a pesar de que el conjunto de datos clasifica esta muestra como carente de cáncer. Sin embargo, es notable que existen áreas dentro de la imagen que el análisis SHAP identifica como características relevantes, coincidiendo con el análisis visual realizado por una experta en la materia. Esta observación sugiere la posible presencia de instancias ruidosas en el conjunto de datos debido a un etiquetado inadecuado. Esta situación resalta la importancia de verificar la precisión de las etiquetas en los datos utilizados para entrenar modelos de aprendizaje automático, especialmente en aplicaciones críticas como el diagnóstico médico.

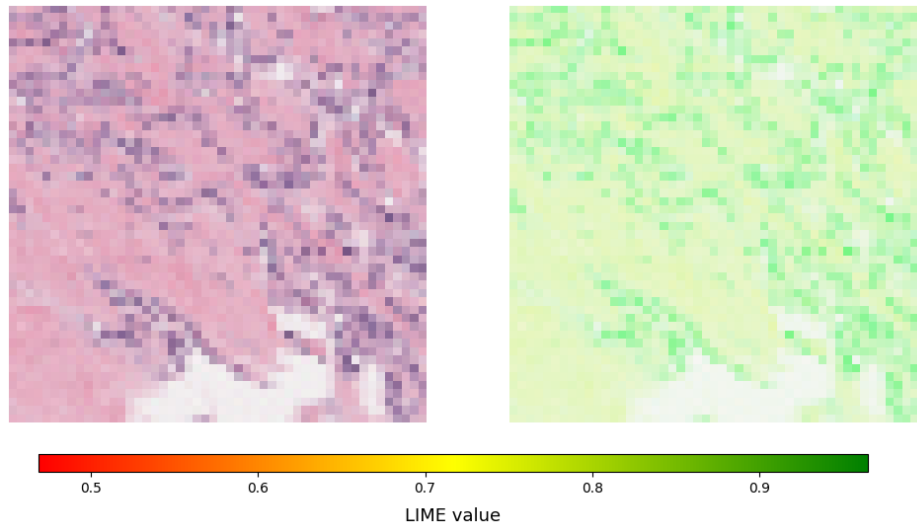


FIGURA 4. *Explicación con LIME para una imagen clasificada erróneamente como sin cáncer, cuando en realidad la etiqueta correcta indica la presencia de cáncer.*

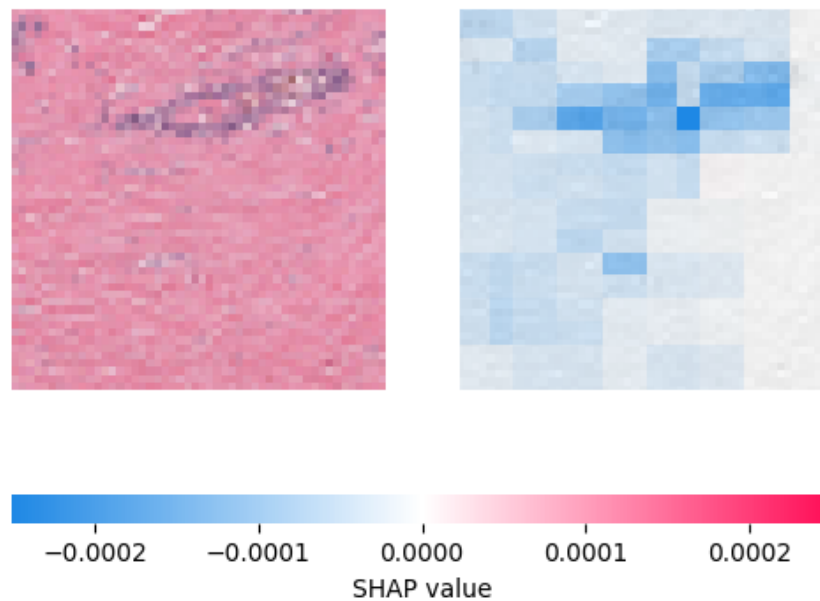


FIGURA 5. *Valores SHAP para una imagen correctamente clasificada como sin cáncer.*

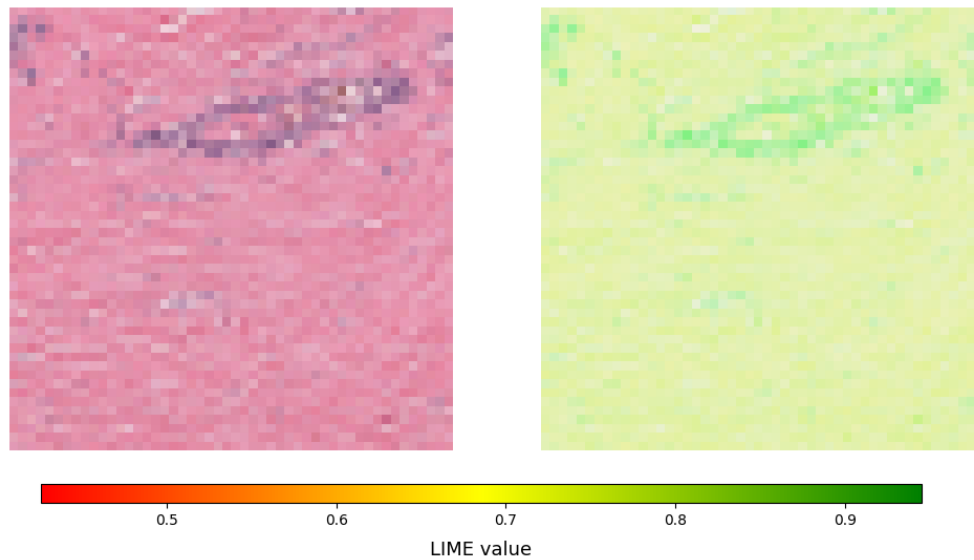


FIGURA 6. Explicación de *LIME* para una imagen correctamente clasificada como sin cáncer.

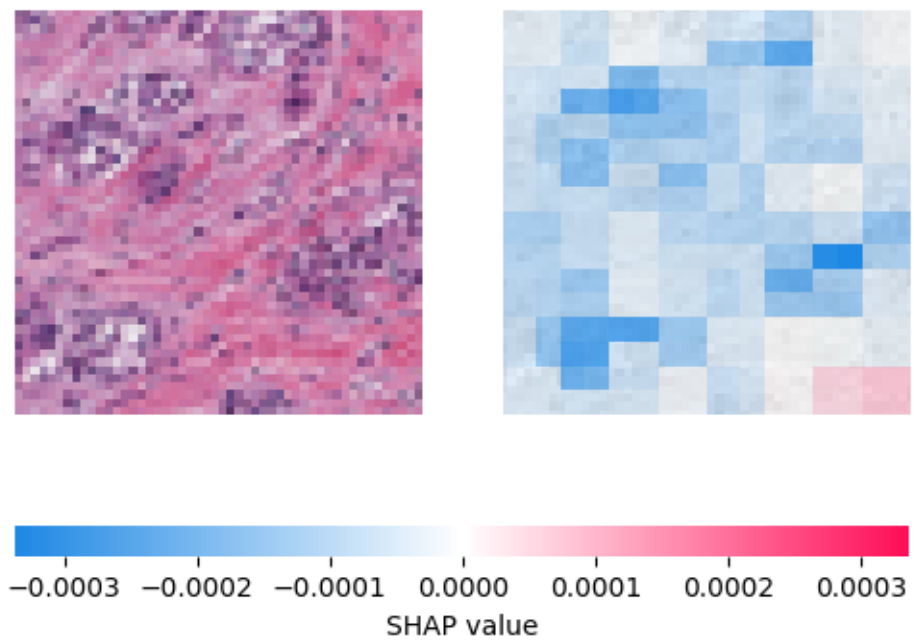


FIGURA 7. Valores *SHAP* para una imagen clasificada erróneamente con presencia de cáncer, cuando en realidad la etiqueta a priori indica ausencia de cáncer.

En la Figura 8, relativa a la aplicación de LIME, se observa un comportamiento similar al descrito por SHAP, donde las zonas marcadas en verde también coinciden con el análisis visual de la experta. Esto refuerza la relevancia de utilizar herramientas de interpretación local, como LIME y SHAP, para validar y entender mejor las predicciones del modelo frente a las etiquetas originales del conjunto de datos.

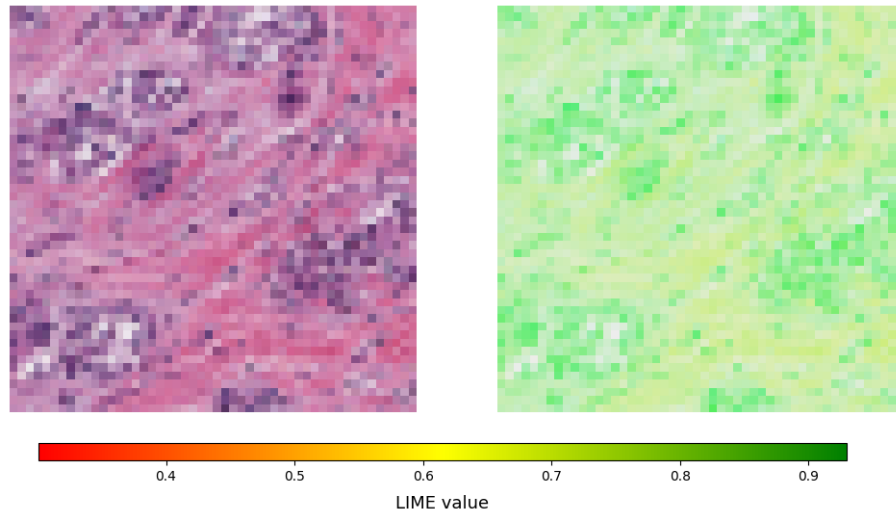


FIGURA 8. *Explicación de LIME para una imagen clasificada erróneamente con presencia de cáncer, cuando en realidad la etiqueta a priori indica ausencia de cáncer.*

Al realizar una comparación entre ambas técnicas de interpretación de modelos, LIME y SHAP, se observa que LIME proporciona una representación detallada, píxel a píxel, de la importancia de las características en la decisión del modelo. Esta granularidad es particularmente útil en el análisis de imágenes de histopatología coloreadas mediante la técnica Hematoxilina y Eosina, ya que permite una elucidación más precisa del aprendizaje del modelo en relación con la base de conocimiento proporcionada.

LIME facilita una mejor comprensión de los patrones específicos y las características morfológicas que el modelo está aprendiendo de las imágenes de histopatología. Esto contribuye a la validación y refinamiento del modelo, asegurando que las inferencias hechas se alineen estrechamente con los criterios diagnósticos establecidos por especialistas médicos.

El conjunto de datos para análisis de cáncer de mama, al centrarse en un subtipo específico, no refleja toda la variedad clínica, lo que lleva a una sobre-optimización del modelo hacia características dominantes no extrapolables a otros contextos. Las técnicas de dropout y regularizaciones implementadas no garantizan eliminar los sesgos hacia características visuales no representativas.

CONCLUSIONES

Este trabajo ha explorado el uso de técnicas de XAI en la interpretación de modelos de IA utilizados para el diagnóstico de cáncer de mama a partir de imágenes histológicas del Carcinoma Ductal Invasivo. A través del uso de SHAP y LIME, se buscó proporcionar transparencia y claridad en las decisiones tomadas por un modelo de red neuronal convolucional entrenado para esta tarea.

Los resultados obtenidos demuestran que ambas técnicas SHAP y LIME, son efectivas para revelar las contribuciones individuales de las características en las predicciones del modelo. El uso de LIME en el análisis de imágenes histopatológicas no solo proporciona una visión integral de las dinámicas de aprendizaje del modelo, sino que también ofrece oportunidades para ajustes dirigidos que pueden mejorar la sensibilidad y especificidad del modelo en tareas de diagnóstico médico.

El uso de XAI en la interpretación de modelos de IA para el diagnóstico de IDC es fundamental para aumentar la confianza en estos sistemas y para facilitar su integración en la práctica médica. La capacidad de explicar y justificar las decisiones del modelo no solo mejora la transparencia, sino que también promueve una mayor aceptación de la IA por parte de los profesionales de la salud.

Como trabajo a futuro se centrará en la mejora del modelo en cuestión de parámetros y el incremento de imágenes en la base de conocimiento del modelo. Adicionalmente, se explorará el impacto que los ajustes en los hiperparámetros pueden tener sobre la precisión y la interpretabilidad del modelo.

Agradecimientos

Angélica Guzmán-Ponce contó con el apoyo del contrato postdoctoral Margarita Salas MGS/2021/23 (UP2021-021) financiado por la Unión Europea-NextGenerationEU

REFERENCIAS

- [1] A. Guzmán-Ponce, J. L. González-Ruíz, and I. Francisco-Valencia. Review on the application of big data algorithms to understand a pandemic virus. In *Handbook of Research on Applied Artificial Intelligence and Robotics for Government Processes*. IGI Global, 2023, pp. 211–233.
- [2] H. Li, N. Zeng, P. Wu, and K. Clawson. Cov-net: A computer-aided diagnosis method for recognizing covid-19 from chest x-ray images via machine vision. *Expert Systems with Applications*, vol. 207, p. 118029, 2022. DOI: 10.1016/j.eswa.2022.118029
- [3] F. Kamalov, K. Rajab, A. K. Cherukuri, A. Elnagar, and M. Safaraliev. Deep learning for covid-19 forecasting: State-of-the-art review. *Neurocomputing*, vol. 511, pp. 142–154, 2022. DOI:

10.1016/j.neucom.2022.09.005

- [4] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. Del Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, and S. Stumpf. Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, vol. 106, p. 102301, 2024. DOI: 10.1016/j.inffus.2024.102301
- [5] L. Malinverno, V. Barros, F. Ghisoni, G. Visonà, R. Kern, P. J. Nickel, B. E. Ventura, I. imi, S. Stryeck, F. Manni, C. Ferri, C. Jean-Quartier, L. Genga, G. Schweikert, M. Lovri, and M. Rosen-Zvi. A historical perspective of biomedical explainable ai research. *Patterns*, vol. 4, no. 9, p. 100830, 2023. DOI: 10.1016/j.patter.2023.100830
- [6] R. O. Alabi, M. Elmusrati, I. Leivo, A. Almangush, and A. A. Mäkitie. Machine learning explainability in nasopharyngeal cancer survival using lime and shap. *Scientific Reports*, vol. 13, no. 1, p. 8984, 2023. DOI: 10.1038/s41598-023-35795-0
- [7] M. M. Abdelsamea, U. Zidan, Z. Senousy, M. M. Gaber, E. A. Rakha, and M. Ilyas. A survey on artificial intelligence in histopathology image analysis. *WIREs Data Mining and Knowledge Discovery*, vol. 12, 2022. DOI: 10.1002/widm.1474
- [8] M. N. S. Choudary, V. B. Bommineni, G. Tarun, G. P. Reddy, and G. Gopakumar. Predicting Covid-19 Positive Cases and Analysis on the Relevance of Features using SHAP (SHapley Additive exPlanation). In *2021 Second ICESC*, Coimbatore, India, 2021, pp. 1892–1896. DOI: 10.1109/ICESC51422.2021.9532829
- [9] K. Borys, Y. A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C. M. Friedrich, and F. Nensa. Explainable ai in medical imaging: An overview for clinical practitioners – saliency-based xai approaches. *European Journal of Radiology*, vol. 162, p. 110787, 2023. DOI: 10.1016/j.ejrad.2023.110787
- [10] M. S. H. Shovon, M. F. Mridha, K. M. Hasib, S. Alfarhood, M. Safran, and D. Che. Addressing uncertainty in imbalanced histopathology image classification of her2 breast cancer: An interpretable ensemble approach with threshold filtered single instance evaluation (sie). *IEEE Access*, vol. 11, pp. 122 238–122 251, 2023. DOI: 10.1109/ACCESS.2023.3327898
- [11] D. Kaplun, A. Krasichkov, P. Chetyrbok, N. Oleinikov, A. Garg, and H. S. Pannu. Cancer cell profiling using image moments and neural networks with model agnostic explainability: A case

study of breast cancer histopathological (breckhis) database. *Mathematics*, vol. 9, no. 20, p. 2616, 2021. DOI: 10.3390/math9202616

- [12] A. Janowczyk and A. Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, vol. 7, no. 1, p. 29, 2016. DOI: 10.4103/2153-3539.186902
- [13] Predicting IDC in Breast Cancer Histology Images. *Kaggle*, 2018, consultado el 29 de mayo del 2024. [en línea]. Disponible: <https://www.kaggle.com/code/paultimothymooney/predicting-idc-in-breast-cancer-histology-images>

Diseño de rutas para estaciones de cargas eléctricas de vehículos eléctricos con algoritmos Metaheurísticos

Design of Routes for Electric Vehicle Charging Stations Using Metaheuristic Algorithms

Ortiz-Aguilar, L.^{1*} , Palacios-Ortega, M.² , Xoca Orozco, L. A.³  (*lucero.oa@purisima.tecnm.mx)

Recibido: 31 de mayo de 2024 Aceptado: 21 de junio de 2024

RESUMEN

En este trabajo se plantea un diseño de rutas para estaciones de carga eléctrica de vehículos eléctricos (VE) en el Estado de Guanajuato, utilizando algoritmos metaheurísticos. El crecimiento de la movilidad eléctrica plantea desafíos significativos en la planificación y optimización de infraestructuras de carga. En respuesta, este artículo propone una metodología basada en algoritmos metaheurísticos, específicamente algoritmos de búsqueda local iterada y la metodología de diseño propuesta por Ortiz-Aguilar et al. (2015) [1], para optimizar la ubicación y distribución de estaciones de carga, maximizando la cobertura de

servicio. Se presenta un estudio de caso detallado, utilizando datos geoespaciales de las estaciones de la República Mexicana, para validar la eficacia y viabilidad de la propuesta. Los resultados muestran que los enfoques heurísticos dentro de un algoritmo de búsqueda local iterada que ofrecen soluciones competitivas en términos de cobertura de servicio, lo que demuestra su utilidad en la planificación de infraestructuras de carga para VE.

Palabras clave: Problema del agente viajero, Aplicaciones de la IA, Metaheurísticas, Vehículos eléctricos, Electromovilidad

¹División de Sistemas Automotrices, Tecnológico Nacional de México / ITS de Purísima del Rincón, Blvd. del Valle 2301, Guardarrayas, 36418 Purísima de Bustos, Gto.

²División de Ingeniería Industrial, Tecnológico Nacional de México / ITS de Purísima del Rincón, Blvd. del Valle 2301, Guardarrayas, 36418 Purísima de Bustos, Gto.

³División de Ingeniería Bioquímica, Tecnológico Nacional de México / ITS de Purísima del Rincón, Blvd. del Valle 2301, Guardarrayas, 36418 Purísima de Bustos, Gto.



ABSTRACT

In this work, a route design for electric vehicle (EV) charging stations in the State of Guanajuato is proposed using metaheuristic algorithms. The growth of electric mobility poses significant challenges in the planning and optimization of charging infrastructures. In response, this article proposes a methodology based on metaheuristic algorithms, specifically Iterated Local Search algorithms and the design methodology proposed by Ortiz-Aguilar [1], to optimize the location and distribution of charging stations, maximizing service coverage. A detailed case study is presented, using geospatial data from stations in the Mexican Republic, to validate the efficacy and feasibility of the proposal. The results show that metaheuristic approaches offer competitive solutions in terms of service coverage, demonstrating their utility in the planning of EV charging infrastructures.

Keywords: *TSP, Metaheuristics, Electric vehicles, Electromobility*

INTRODUCCIÓN

En el contexto actual de los vehículos se está optando por cambiar de los automóviles de combustión interna a los eléctricos. En México se busca reducir las emisiones de gases de efecto invernadero, ya que de acuerdo con el inventario Nacional de Emisiones (1990-2019) [2], menciona que el 24.5 % de las emisiones de CO_2 vienen del sector transporte. El nuevo reto será el generar la infraestructura necesaria para ubicar estaciones de carga para vehículos eléctricos. Al igual que las gasolineras tradicionales, las estaciones de carga pueden tener diferentes capacidades y opciones de precios y el costo de recarga de cada opción puede variar de una estación a otra.

Una estrategia de carga inteligente puede contribuir de manera importante a la gestión eficiente de los recursos disponibles en las redes de los Vehículos eléctricos (EV). Al reconocer la importancia de la carga inteligente en el contexto de las redes EV, algunas investigaciones han presentado varias estrategias de carga inteligente en los últimos años, cuyo objetivo es aumentar la autonomía y el tiempo de carga [3].

Dado el contexto anterior, se puede considerar para calcular el costo de energía el camino hasta el destino para reducir el tiempo y el coste de la carga. Esto proporcionó la motivación para esta investigación.

La propuesta de este trabajo es considerar las estaciones de gasolina existentes como posibles puntos de recarga para los EV, dado que la demanda de combustibles disminuirá a medida que aumente el número de vehículos eléctricos. Para resolver este problema, se utilizó una técnica del estado del arte bien conocida como el algoritmo de búsqueda local iterada para aproximar y ofrecer una solución factible.

El artículo se organiza de la siguiente manera, en la sección Estado del Arte se describe el estado

del arte de la investigación. En la sección Marco Teórico, se describe el marco teórico y en la sección Metodología la metodología. En la sección Resultados se reportan los resultados de esta investigación y finalmente en la última sección se dan conclusiones del tema.

ESTADO DEL ARTE EN ELECTROMOVILIDAD Y SUS APLICACIONES CON IA

En esta sección se da una visión general de estado del arte en las investigaciones relacionadas a la electromovilidad y aplicaciones de la inteligencia artificial a vehículos eléctricos.

En la investigación realizada por Rajawat et al. (2024) [4] estudian el problema de coordinación de carga, donde un conjunto de camiones tiene como objetivo planificar dónde y durante cuánto tiempo cargar a lo largo de rutas previamente planificadas para su propio beneficio, mientras que los recursos de carga son limitados. Para facilitar la planificación de la carga de camiones y mitigar la congestión en las estaciones de carga, dicho trabajo propone un marco de coordinación distribuida de dos capas. Los autores calculan los planes de carga de los camiones resolviendo problemas de optimización de enteros mixtos, y las estaciones programan las órdenes de carga de acuerdo con la regla de "*primero en llegar, primero en ser atendido*". Su esquema solo requiere un simple intercambio de información entre camiones y estaciones, esto permite cálculos completamente distribuidos y la replanificación por parte de los camiones al acercarse a cada estación. Además, realizaron estudios numéricos basados en la red de carreteras sueca y compararon los resultados con planes de carga calculados fuera de línea. Se demostró que el marco trajo grandes beneficios al reducir los costos colectivos de todas las partes involucradas. Además, proponen integrar la estimación de multitudes en tiempo real en las estaciones de carga.

En el trabajo realizado por Silber et al. (2024) [5], se demostró el potencial para reducir los factores de simultaneidad de los sitios de carga rápida. Adoptarlo ayudaría en la expansión de la infraestructura de carga de vehículos eléctricos, ya que, por un lado, significa que los transformadores pueden ser más pequeños y, consecuentemente, la demanda de energía de los operadores de red en los puntos de conexión puede ser menor. Esto ayuda en la transmisión de energía y no reserva una cantidad máxima de energía raramente necesaria, lo que ralentizaría la construcción de otros sitios de carga. Por otro lado, los operadores de puntos de carga pueden ahorrar recursos al diseñar sitios de carga rápida y, por lo tanto, pueden construir más. Esto aumentaría las oportunidades de carga para los conductores y ayudaría a la aceptación de la electromovilidad en general. Sin embargo, el umbral de aceptación α se estableció en 0.01, lo que influye en los resultados de *OPT*. En investigaciones futuras, se deberían investigar los valores aceptables para α para posiblemente aumentar el umbral y disminuir aún más *OPT*. Además, los avances tecnológicos en el diseño de estaciones de carga podrían integrarse, así como mejoras en las curvas de carga de vehículos eléctricos. Esto revelaría qué tan estables resultan ser los resultados de este trabajo ante la tecnología cambiante. Por último, el análisis sobre la integración de sistemas de almacenamiento de energía eléctrica (BESS) puede llevarse a cabo con más detalle dadas los resultados

negativos de esta publicación. Cambiar las tasas de carga y las capacidades, así como reducir los precios, por ejemplo, para el uso en segunda vida, puede llevar a soluciones económicamente viables.

Por otra parte, en la investigación realizada por Chen, et al. (2023) [6], se establece un marco de coordinación P2P jerárquico para estaciones de carga de vehículos eléctricos (EVCS) equipadas con tecnología de respuesta a la demanda (DRE) y sistemas de almacenamiento de energía eléctrica (BESS), donde la transacción P2P entre múltiples EVCSs se coordina con la operación de redes de transporte urbano (UTN) y la operación de distribución de energía (PDN). Se desarrolla un esquema de transacción P2P completamente descentralizado para múltiples EVCSs, que también incorpora la protección de la privacidad y el costo comercial de las transacciones de energía (TE). Los resultados muestran que, primero, los resultados de asignación de carga de los vehículos eléctricos (EVs) pueden estar fácilmente disponibles para cada agente EVCS después de resolver el problema de asignación de transporte de energía múltiple (MTAP) por el operador del sistema de transporte (TSO). La plataforma P2P permite transacciones de TE entre EVCSs, donde la compra de energía de la red se reduce en casi un 25.44 %, de 6161.7 kWh a 4,594 kWh. La energía intercambiada entre EVCSs en el proceso de negociación P2P es de hasta 2,583.7 kWh, lo que permite una reducción del 16.61 % en el costo de operación de EVCS en comparación con el funcionamiento independiente. Finalmente, se señala que algunos casos valiosos indican que la reconfiguración de la PDN puede mejorar el espacio factible de transacciones P2P entre EVCSs bajo las restricciones de operación del DSO. Además, se investiga el impacto de la congestión del tráfico de UTN y se valida la escalabilidad del marco propuesto. En el sistema de 123 buses de IEEE, donde hay más entidades comerciales y mayores cantidades de transacción, se logra una reducción del 39.4 % en el costo de operación de EVCS.

En el trabajo realizado por Moghaddam et al. (2017) [7], proponen una estrategia de carga inteligente para una red de vehículos eléctricos enchufables, la cual ofrece múltiples opciones de carga en las estaciones de carga. Consideran que las diferentes restricciones presentadas como lo pueden ser la distancia, tiempo, etc., son importantes de considerar para que el usuario del vehículo pueda recargar en un tiempo mínimo y llegar a su destino sin retrasos significativos. Como método de solución utilizan un método de colas para estimar el retraso en varias estaciones de carga y para evitar los tiempos de espera largos y la posible superposición entre periodos pico de carga introdujeron el concepto de carga parcial. Los resultados reportados muestran que en la simulación reducen significativamente el retraso medio en la carga (hasta un 25 %) y el costo (hasta un 15 %). Como trabajo futuro proponen investigar con un modelo de precios dinámico óptimo para minimizar la superposición entre el PEV pico y el residencial.

En el contexto del diseño de rutas para estaciones de carga de vehículos eléctricos (EV), la planificación y optimización de infraestructuras de carga es un desafío complejo que implica múltiples variables y restricciones. Las metaheurísticas ofrecen una serie de ventajas que las hacen particularmente adecuadas

para abordar este problema.

MARCO TEÓRICO

Electric Vehicle Charging Scheduling

El problema de la programación de carga de vehículos eléctricos o Electric Vehicle Charging Scheduling (EVCS), que tiene como objetivo determinar es encontrar una estación de carga a lo largo del camino de modo que el tiempo total de viaje (incluido el tiempo de conducción desde la ubicación actual hasta el destino, el tiempo de espera en una estación de carga y el tiempo de carga) y el costo de carga sean minimizado [7]. De Acuerdo con Moghaddam et al. (2017) [7], este problema se formula de acuerdo con Ec. (1):

$$T^{\{drive\}}(s, dst) = \frac{D(s, j)}{v(s, j)} + \frac{D(j, dst)}{v(j, dst)} \quad (1)$$

donde:

- $D(s, j)$ y $D(j, dst)$ indican la distancia de la ruta desde el origen hasta la estación de carga y desde la estación de carga hasta el destino, respectivamente.
- $v(s, j)$ y $v(j, dst)$ indican la velocidad de movimiento promedio del PEV desde su origen hasta la estación de carga y desde la estación de carga hasta el destino, respectivamente.

El costo de carga de cada vehículo eléctrico enchufable (PEV) depende de su opción de carga asociada y de la tarifa de carga asociada en la estación j . Para un PEV, el costo de carga se calcula mediante la Ec. (2):

$$C_j^{recharge}(o) = (d_{charging} Pr_k) j \text{ en } j \quad (2)$$

donde $d_{charging}$ es la demanda de carga de cada PEV, que depende del SoC restante y objetivo para cada PEV (Ec. (3)):

$$d_{charging} = \sigma E - SoC \quad (3)$$

Aquí, σ es el coeficiente de carga parcial, que puede variar desde el SoC actual hasta el SoC máximo (por ejemplo, 90 % de la capacidad de la batería). En circunstancias normales, la mayoría de los usuarios preferirían una carga completa (es decir, $\sigma = 0,9$).

Heurísticas de perturbación

Una definición simple dada por Singh y Choudhary (2021) [8] es "*El enfoque heurístico se basa en prueba y error para encontrar o descubrir soluciones a los problemas*". En metaheurística se utilizan con mayor frecuencia dos tipos de heurísticas: construcción y perturbación. De acuerdo con lo mencionado anteriormente, se definen las heurísticas constructivas y perturbativas utilizadas en este trabajo. Según Laporte y F. Semet [9] dos técnicas principales para construir soluciones al problema del viajante (TSP) son la heurística codiciosa o aleatoria y la definición formal de TSP se puede encontrar en lo publicado por Jünger et al. (1995) [10].

En esta investigación se eligió la heurística aleatoria para realizar las soluciones iniciales (ver Algoritmo 1) y cinco heurísticas perturbativas descritas en los Algoritmos 1, 2, 3, 4, y 5.

Algoritmo 1: Algoritmo de construcción de solución de TSP

Require: $n \geq 0$, $FF \leftarrow \text{Fitness Function}$

1: $S \leftarrow \text{empty solution}$

2: $\text{route} \leftarrow \text{empty list of size } n$

3: **Hacer desde** $i \leftarrow 0$ **hasta** $n - 1$ **hacer**

4: $\text{route}[i] \leftarrow \text{random value between } 0 \text{ and } n - 1$

5: **Fin**

6: $S \leftarrow \text{set}(\text{route})$

7: $S \leftarrow \text{evaluateSolution}(S, FF)$

8: **Regresar** S

Algoritmo 2: Heurística Two Swap (h_1)

Requiere: $n \geq 1$, $SI \leftarrow \text{Initial Solution}$, $FF \leftarrow \text{Fitness Function}$

1: $\text{solutionB} \leftarrow \text{copy}(SI)$

2: $\text{index} \leftarrow \text{random}(0, n - 1)$

3: $\text{index2} \leftarrow \text{random}(0, n - 1)$

4: $\text{value} \leftarrow \text{solutionB}[\text{index2}]$

5: $\text{solutionB}[\text{index2}] \leftarrow \text{solutionB}[\text{index}]$

6: $\text{solutionB}[\text{index}] \leftarrow \text{value}$

7: $\text{solutionB} \leftarrow \text{evaluateSolution}(\text{solutionB}, FF)$

8: **Regresar** $\text{criteria.toReplace}(SI, \text{solutionB})$

Algoritmo 3: heurística Crossover (h_2)

Requiere: $n \geq 1$, $solutionA \leftarrow Initial\ Solution$, $FF \leftarrow Fitness\ Function$

```
1:  $solutionB \leftarrow copy(solutionA)$ 
2:  $parent1 \leftarrow solutionA.body$ 
3:  $n \leftarrow len(parent1)$ 
4:  $parent2 \leftarrow sample(range(0, n), n)$ 
5:  $k \leftarrow random(0, n - 1)$ 
6:  $Index \leftarrow []$ 
7: For{ $i$  in  $range(n)$ }
8: If{ $i \neq k$ }
9:  $Index.append(i)$ 
10: EndIf
11: EndFor
12:  $son \leftarrow []$ 
13: For{ $i$  in  $range(n)$ }
14: If{ $i \geq n - len(Index)$ }
15:  $son.append(parent1[Index[i]])$ 
16: Else
17:  $son.append(-1)$ 
18: EndIf
19: EndFor
20:  $if \leftarrow []$ 
21: For{ $elem$  in  $parent2$ }
22: If{ $elem$  not in  $son$ }
23:  $f.append(elem)$ 
24: EndIf
25: EndFor
26: For{ $i$  in  $range(n)$ }
27: If{ $son[i] == -1$ }
28:  $son[i] \leftarrow f.pop(0)$ 
29: EndIf
30: EndFor
31:  $solutionB \leftarrow set(son)$ 
32:  $solutionB \leftarrow evaluateSolution(solutionB, FF)$ 
33: Regresar  $criteria.toReplace(SI, solutionB)$ 
```

Algoritmo 4: Heurística Nearest Neighbor (h_3)

Requiere: $n \geq 1$, $solutionA \leftarrow Initial\ Solution$, $FF \leftarrow Fitness\ Function\ matrix$ $cost\ Matrix$

```

1:  $solutionB \leftarrow copy.deepcopy(solutionA)$ 
2:  $parent1 \leftarrow solutionA.body$ 
3:  $c \leftarrow random.randint(0, len(parent1) - 1)$ 
4:  $ic \leftarrow parent1.index(c)$ 
5:  $rowt \leftarrow matrix[c]$ 
6:  $rowOrderedWithIndex \leftarrow []$ 
7: For{ $i$ ,  $value$  in  $enumerate(rowt)$ }
8:  $rowOrderedWithIndex.append((i, value))$ 
9: EndFor
10:  $rowOrderedWithIndex.sort(rowt)$ 
11:  $orderedIndex \leftarrow []$ 
12: For{ $index$ , in  $rowOrderedWithIndex$ }
13:  $orderedIndex.append(index)$ 
14: EndFor
15:  $nc \leftarrow orderedIndex[1]$ 
16: For{ $i$ , in  $enumerate(parent1)$ }
17: If{ $parent1[i] == nc$ }
18:  $inc \leftarrow i$ 
19: break
20: EndIf
21: EndFor
22:  $c1 \leftarrow parent1[ic + 1]$ 
23:  $parent1[ic + 1] \leftarrow nc$ 
24:  $parent1[inc] \leftarrow c1$ 
25:  $solutionB \leftarrow set(son)$ 
26:  $solutionB \leftarrow evaluateSolution(solutionB, FF)$ 
27: Regresar  $criteria.toReplace(SI, solutionB)$ 

```

Algoritmo 5: Heurística K-OPT (h_4 , h_5)

Requiere: $n \geq 1$, $solutionA \leftarrow Initial\ Solution$, $FF \leftarrow Fitness\ Function$

```
1:  $solutionB \leftarrow copy(solutionA)$ 
2:  $son \leftarrow solutionB.body$ 
3:  $ind \leftarrow random.sample(len(son), n)$ 
4:  $random.shuffle(ind)$ 
5:  $sonCopy \leftarrow son[:]$ 
6: For{ $i$  in  $range(self.n)$ }
7:  $son[ind[i]] \leftarrow sonCopy[ind[(i + 1) \% n]]$ 
8: EndFor
9:  $solutionB \leftarrow set(son)$ 
10:  $solutionB \leftarrow evaluateSolution(solutionB, FF)$ 
11: Regresar  $criteria.toReplace(SI, solutionB)$ 
```

Metaheurística de Búsqueda Local Iterada

En la experimentación se utilizó el algoritmo de búsqueda local iterada (ILS). Esta metaheurística fue propuesta por Lourenco et al. (2023) [11], que construyen o modifican una solución mediante una heurística incorporada. Las soluciones generadas son mejores que las generadas aleatoriamente o alteradas. Este algoritmo tiene como objetivo intensificar la solución inicial explorando sus vecinos. El algoritmo ILS se muestra en el Algoritmo 6, tomado de Talbi (1965) [12].

METODOLOGÍA

La metodología de este trabajo se basa en la propuesta por Ortiz-Aguilar et al. [13] y se puede resumir en los siguientes pasos o etapas, cada uno de los cuales se describe detalladamente a continuación:

I. Diseño del insumo

■ Modelado de Restricciones y Variables:

Matriz de Distancias (MMA): Se recopiló la ubicación geográfica de las estaciones de gasolina existentes en el estado de Guanajuato. Utilizando Python y Google Maps, se obtuvo la longitud y la latitud de cada estación para generar una matriz de distancias entre todas las estaciones. Esta matriz es fundamental para calcular las rutas óptimas.

Parámetros de Carga: Se consideraron variables como la capacidad de carga de las estaciones, las tarifas de carga, y las restricciones de tiempo. Estas variables se integraron en el modelo para asegurar que las soluciones sean realistas y aplicables en un contexto práctico.

Algoritmo 6: Iterated Local Search (ILS)

Requiere: $inSol \leftarrow Initial\ Solution$, $maxI \leftarrow MaxIterations$

```

1:  $bestSol \leftarrow inSol$ 
2:  $currentSol \leftarrow inSol$ 
3:  $iteration \leftarrow 0$ 
4: While{ $iteration < maxI$ }
5:  $perturbedS \leftarrow perturb(currentSol)$ 
6:  $localOptima \leftarrow localSearch(perturbedSol)$ 
7: If{ $isImprovement(localOptima, bestSol)$ }
8:  $bestSol \leftarrow localOptima$ 
9: EndIf
10:  $currentSol \leftarrow diversify(localOptima)$ 
11:  $iteration \leftarrow iteration + 1$ 
12: EndWhile
13: Regresar  $bestSol$ 

```

II. Diseño Metodológico de los Experimentos

Selección de técnicas metaheurísticas y parámetros

- **Metaheurísticas Utilizadas:** Se seleccionó la búsqueda local iterada (ILS) para resolver el problema. La elección de esta técnica se basa en su capacidad comprobada para manejar problemas de optimización complejos y su flexibilidad para adaptarse a diferentes condiciones.

- **Heurísticas Constructivas y Perturbativas:** Para generar soluciones iniciales y mejorar las existentes, se emplearon varias heurísticas, como:

Heurística Aleatoria: Para generar soluciones iniciales aleatorias.

Heurística Two-Swap : Intercambio de dos elementos en la solución.

- **Heurística Crossover:** Combinación de soluciones parentales para generar nuevas soluciones.

- Heurística Nearest Neighbor: Selección del vecino más cercano para construir rutas.
- Heurística K-OPT: Reordenamiento de sub-rutas para mejorar la solución.

III. Experimentación

Configuración y ejecución de experimentos

La implementación se realizó utilizando Python y las condiciones de experimento. Cada algoritmo fue probado bajo condiciones controladas con un mismo límite de llamadas a función por instancia. Los experimentos se repitieron 31 veces para asegurar la consistencia y fiabilidad de los resultados.

Estadísticas Utilizadas: Para el análisis estadístico de los resultados, se aplicaron pruebas como Shapiro-Wilks para verificar la normalidad de los datos. Dado que los datos no seguían una distribución normal, se utilizó la mediana como medida representativa. Además, se emplearon pruebas de Friedman, Friedman Alineado y Quade para comparar el rendimiento de las heurísticas.

IV. Evaluación de resultados

Análisis de Resultados y Comparación

- Métricas de Desempeño: Los resultados se evaluaron en términos de distancia total recorrida y la eficiencia de las rutas generadas. Se compararon las heurísticas utilizando métricas de desempeño específicas para cada técnica metaheurística.
- Selección de Mejores Soluciones: Se identificaron las dos mejores heurísticas basadas en su desempeño.

RESULTADOS

Para este conjunto de casos, el enfoque se inclinó hacia el empleo de las estructuras *MMA* y *LPH*. Estos fueron seleccionados en función de su relevancia y aplicabilidad en el marco del problema, como se detalla en el trabajo Ortiz-Aguilar et al. (2023) [14].

A continuación, se describe la configuración de parámetros para cada Metaheurística y se explica las pruebas estadísticas utilizadas para comparar los experimentos. La experimentación implicó probar cada una de las instancias en una computadora portátil con las siguientes especificaciones: una CPU Intel® Core i3-1005G1 con frecuencia de 1,20 GHz, 8 GB de RAM y un sistema operativo Windows 11 de 64 bits. La metodología se implementó utilizando Python y los experimentos se repitieron 31 veces. Cada experimento heurístico tenía un límite estipulado de 100.000 llamadas a funciones por instancia.

Para el análisis estadístico, primero se aplicó la prueba estadística de Shapiro-Wilks para comprobar si los datos tenían una distribución normal, y los resultados indicaron que los datos no seguían una distribución normal; por lo tanto, se eligió la mediana como estadísticamente representativa.

Instancias

Como punto de partida para hacer una propuesta de solución de dónde ubicar posibles puntos de carga de vehículos eléctricos fue tomar la información de donde se encuentran actualmente las gasolineras en México. Por lo cual se consultó el informe de estaciones de servicio y estaciones de autoconsumo franquiciadas por Petróleos Mexicanos (PEMEX) obtenida del Directorio de Estaciones de Servicio [15], en la Tabla 1 se muestra el resumen de gasolineras por estado que están activas, siendo un total de 6754.

De las instancias para este trabajo se decidió tomar primero las del estado de Guanajuato para hacer el análisis. La información obtenida contenía solamente la dirección de cada estación, por lo que era necesario sacar una coordenada en términos de longitud y latitud para poder procesarla y generar una matriz de costos (MMA).

Tabla 1. Cantidad de estaciones de gasolina por Estado en México.

Estado	No	Estado	No
Campeche	77	Oaxaca	111
Chiapas	155	Puebla	286
Chihuahua	215	Querétaro	196
Ciudad De México	297	Quintana Roo	79
Coahuila	237	San Luis Potosí	148
Colima	45	Sinaloa	220

Tabla 1. Cantidad de estaciones de gasolina por Estado en México (...continuada).

Estado	No	Estado	No
Durango	93	Sonora	248
Estado De México	670	Tabasco	169
Guanajuato	434	Tamaulipas	218
Guerrero	90	Tlaxcala	54
Hidalgo	148	Veracruz	448
Jalisco	461	Yucatán	177
Michoacán	187	Zacatecas	85

En una primera aproximación se supone que se tienen que pasar por todas las estaciones para tener un tiempo de recarga estándar en cada una de ellas, por lo cual a este punto es solamente importante la distancia total recorrida. En la Tabla 2 se muestran los resultados en términos de distancia de las heurísticas y se tomarán a las dos mejores para ser puestas en el algoritmo de búsqueda local con las diferentes heurísticas.

Tabla 2. Resultados de las heurísticas.

Estado	h_1	h_2	h_3	h_4	h_5
AGUASCALIENTES	2.89	6.61	4.53	2.91	3.14
BAJA CALIFORNIA	58.83	240.63	154.65	54.79	67.92
BAJA CALIFORNIA SUR	30.29	82.95	50.69	24.55	23.14
CAMPECHE	16.08	58.20	32.41	18.38	16.00
CHIAPAS	34.37	146.32	93.64	38.51	44.64
CHIHUAHUA	73.09	306.76	205.33	87.18	95.65
COAHUILA	110.34	403.59	228.26	107.18	108.10
COLIMA	4.10	8.05	4.85	4.01	1.91
DURANGO	19.25	75.47	47.73	22.22	24.01
ESTADO DE MEXICO	78.67	293.35	196.33	99.81	119.00
GUERRERO	27.93	76.76	41.14	22.86	19.61
HIDALGO	26.26	71.62	45.93	23.63	24.47
JALISCO	93.74	372.71	242.09	117.48	143.48
MICHOACAN	45.78	170.11	105.03	55.40	55.63

Tabla 2. Resultados de las heurísticas (...continuada).

Estado	h_1	h_2	h_3	h_4	h_5
MORELOS	6.45	16.81	9.81	5.60	7.40
NAYARIT	45.53	68.78	56.93	36.06	35.17
NUEVO LEON	63.16	170.00	117.33	71.00	78.19
OAXACA	39.72	134.82	78.45	40.42	40.00
PUEBLA	39.54	144.52	91.72	48.21	54.09
QUERETARO	15.24	52.76	37.15	20.01	21.43

Por último, se aplicó una prueba de Friedman, Friedman Alineado y Quade, las cuales se utilizan frecuentemente para comparar algoritmos en Joaquín Derrac et al. [16], de dichas pruebas se muestran los resultados de la Tabla 3, de la cual se puede deducir lo siguiente:

- h_1 : Tiene valores de 1.75 en Friedman y Quade, y 596 en Friedman Alineado, sugiriendo un rendimiento relativamente bajo en comparación con otras heurísticas.
- h_2 : Con valores de 5, 1732, y 5, respectivamente, esta heurística muestra un rendimiento significativamente diferente, probablemente peor o mejor que las otras dependiendo del contexto del experimento.
- h_3 : Valores de 4, 1488, y 4, respectivamente, indicando un rendimiento intermedio.
- h_4 : Con valores de 1.9, 600, y 1.9, muestra un rendimiento similar al de h_1 .
- h_5 : Valores de 2.35, 634, y 2.35, también sugieren un rendimiento cercano al de h_1 y h_4 .

Tabla 3. Estadísticos de las pruebas de Friedman, Friedman Alineado y Quade.

Prueba	h_1	h_2	h_3	h_4	h_5	Estadístico
Friedman	1.75	5	4	1.9	2.35	65.56
Friedman Alineado	596	1732	1488	600	634	62.1871415
Quade	1.75	5	4	1.9	2.35	30.7795386

La mejor heurística de acuerdo con las pruebas, el iterado con los valores más bajos es la heurística de exploración h_1 y, por lo tanto, es el que tiene mejor rendimiento.

CONCLUSIONES

Este estudio presenta una comparativa de diversas heurísticas aplicadas al diseño de rutas para estaciones de carga de vehículos eléctricos. Los resultados obtenidos demuestran que los enfoques metaheurísticos, como la búsqueda local iterada, ofrece una solución inicial al dicho problema. Sin embargo, es importante destacar que esta investigación se enfoca exclusivamente en la comparación de diferentes heurísticas y sus aplicaciones, sin pretender establecer una solución definitiva al problema planteado.

En resumen, aunque esta investigación proporciona una base sólida para el diseño de rutas para estaciones de carga de vehículos eléctricos utilizando algoritmos metaheurísticos, es evidente que se necesita un enfoque multidisciplinario y colaborativo para abordar las limitaciones y desafíos identificados. Futuros trabajos deben enfocarse en mejorar la precisión de los modelos, integrar tecnologías avanzadas, y considerar más profundamente los factores humanos y económicos para optimizar la infraestructura de carga de EV de manera holística.

REFERENCIAS

- [1] L. de M. O. Aguilar, J. M. C. Valadez, H. J. P. Soberanes, C. L. D. González, C. L. Ramirez, and J. Soria-Alcaraz. Comparativa de algoritmos bioinspirados aplicados al problema de calendarización de horarios. *Research in Computing Science*, (94), pp. 33–43, 2015.
- [2] Inventario Nacional de Emisiones de Gases y Compuestos de Efecto Invernadero, 1990-2019.
- [3] N. Rauh, T. Franke, and J. F. Krems. Understanding the impact of electric vehicle driving experience on range anxiety. *Hum Factors*, vol. 57, no. 1, pp. 177–187, 2015.
- [4] S. P. Rajawat, A. Gautam, S. J. Janarthanan, L. Soni, and P. Khandekar. Adaptive EV Range Estimation and Optimization Based on Rider Demand and Terrain Requirements, 2024.
- [5] F. Silber, S. Scheubner, and A. März. Analysis of the simultaneity factor of fast-charging sites using Monte-Carlo simulation. *International Journal of Electrical Power & Energy Systems*, vol. 155, p. 109540, 2024.
- [6] X. Chen, X. Wang, M. Shahidehpour, L. Affolabi, Z. Lu, and K. Li. Distributed Peer-to-Peer Coordination of Hierarchical Three-Phase Energy Transactions Among Electric Vehicle Charging Stations in Constrained Power Distribution and Urban Transportation Networks. *IEEE Transactions on Transportation Electrification*, 2023.

- [7] Z. Moghaddam, I. Ahmad, D. Habibi, and Q. V. Phung. Smart charging strategy for electric vehicle charging stations. *IEEE Transactions on Transportation Electrification*, vol. 4, no. 1, pp. 76–88, 2017.
- [8] S. K. Singh Padam and Choudhary. Introduction: Optimization and Metaheuristics Algorithms. In *Metaheuristic and Evolutionary Computation: Algorithms and Applications*, A. and J. P. and A. S. and B. F. I. Malik Hasmat and Iqbal, Ed., Singapore: Springer Singapore, 2021, pp. 3–33.
- [9] G. Laporte and F. Semet. Classical heuristics for the capacitated VRP. In *The vehicle routing problem*, SIAM, 2002, pp. 109–128.
- [10] M. Jünger, G. Reinelt, and G. Rinaldi. The traveling salesman problem. *Handbooks in operations research and management science*, vol. 7, pp. 225–330, 1995.
- [11] L. Helena, M. Olivier, and S. Thomas. Iterated Local Search. In *Handbook of Metaheuristics*, vol. 57, F. Glover, G. Kochenberger, and F. S. Hillier, Eds., International Series in Operations Research & Management Science, vol. 57. Springer New York, 2003, pp. 320–353.
- [12] E. G. Talbi. *Metaheuristic from Design to Implementation*. Wiley, 1965.
- [13] L. de M. Ortiz-Aguilar, M. Carpio, H. Puga, J. A. Soria-Alcaraz, M. Ornelas-Rodríguez, and C. Lino. Increase methodology of design of course timetabling problem for students, classrooms, and teachers. In *Nature-Inspired Design of Hybrid Intelligent Systems*, pp. 713–728, 2017.
- [14] L. Ortiz-Aguilar, H.-A. Yeovanna, M. Benitez, S. Rodriguez-Miranda, and F. Mendoza-Vazquez. A Comparison Between Selection Operators Heuristics of Perturbation in CSP. In *Hybrid Intelligent Systems Based on Extensions of Fuzzy Logic, Neural Networks and Metaheuristics*, Springer Nature Switzerland Cham, 2023, pp. 365–377.
- [15] Directorio de Estaciones de Servicio. Consultado: el 19 de junio de 2024. [En línea]. Disponible en: https://491231.fs1.hubspotusercontent-na1.net/hubfs/491231/site_edenred_mx/wp/Directorio-Estaciones-Servicio-TC-0722.pdf
- [16] S. G. Joaquín Derrac. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm Evol Comput*, 2011.

Evaluación Sísmica del Tren Interurbano México-Toluca con simulaciones de carga viva y Ciencia de Datos

Seismic Evaluation of the Mexico-Toluca Interurban Train with Live Load Simulations and Data Science

Valdovinos-Rosas, R.M.^{1*} , Arellano-Magdaleno, O. D.¹ , López-Díaz, J. L.¹ , Delgado-Hernández, D. J.¹ 
(*rvaldovinosr@uaemex.mx)

Recibido: 6 de junio de 2024 Aceptado: 19 de junio de 2024

RESUMEN

El Gobierno Federal de México está construyendo el Tren Interurbano México-Toluca para abordar el desafío de transportar a más de 230 mil personas diariamente entre las ciudades de México y Toluca. Para evaluar la construcción del tren se han propuesto estudios, de los que la mayoría se realizan a base de simulaciones por la poca disposición de datos reales. En este artículo se realiza el análisis del modelo estructural del tren interurbano México-Toluca al ser sometido a un terremoto de magnitud similar al del 22 de mayo de 1960 en Chile y del 19 de septiembre de 1985 en México considerando simulaciones ejecutadas

con el programa SAP 2000, mientras el tren se encuentra en movimiento a una velocidad determinada y con carga máxima de personas. Para la investigación se utiliza el lenguaje de programación Python y el software libre WEKA siguiendo el proceso general de ciencia y analítica de datos. El análisis de clúster mostró un severo traslape de clases en los datos del sismo en México. La clasificación mejoró significativamente al usar conjuntos derivados del agrupamiento con el algoritmo EM.

Palabras clave: *Tren interurbano, Sismo, Ciencia de datos, Agrupamiento, Machine Learning.*

¹Facultad de Ingeniería de la Universidad Autónoma del Estado de México. Paseo Universidad S/N. Toluca, Estado de México, México. CP. 50130.



ABSTRACT

The Federal Government of Mexico is constructing the Mexico-Toluca Interurban Train to address the challenge of transporting over 230,000 people daily between the cities of Mexico and Toluca. To evaluate the train's construction, various studies have been proposed, most of which rely on simulations due to the limited availability of real data. This article analyzes the structural model of the Mexico-Toluca interurban train when subjected to an earthquake of similar magnitude to the ones on May 22, 1960, in Chile and September 19, 1985, in Mexico. The simulations are executed using the SAP 2000 program while the train is in motion at a specific speed and with maximum passenger load. The research utilizes Python programming language and the open-source software WEKA, following the general data science and analytics process. The cluster analysis revealed a severe overlap of classes in the Mexico earthquake data. Classification significantly improved when using sets derived from the clustering with the EM algorithm.

Keywords: *Interurban train, Earthquake, Data science, Clustering, Machine learning.*

INTRODUCCIÓN

El Tren Interurbano México-Toluca es un proyecto del Gobierno Federal que busca conectar la capital del país con la ciudad de Toluca en un tiempo de viaje de 40 minutos. Este medio de transporte tendrá una longitud total de 57.7 km y contará con 7 estaciones (2 terminales y 5 intermedias) [1]. El análisis de viaductos elevados y puentes es importante, ya que pueden afectar estructuras como terremotos, sobrecargas, lluvia en exceso, falta de mantenimiento, plasticidad, fatiga, fractura, etc. Estas situaciones pueden provocar consecuencias negativas en las estructuras y desencadenar eventos no deseados, por ello, para evaluar su seguridad se analiza la estructura de la obra.

México, está ubicado en la placa tectónica norteamericana y limitando al sur y oeste con las placas de Cocos, Rivera y del Pacífico, se encuentra en el “Cinturón de Fuego”, una zona sísmicamente activa a nivel mundial. Esta situación motiva el estudio presentado en este artículo, evaluando la respuesta del tren Interurbano ante un terremoto mientras está en movimiento y transportando el máximo número de pasajeros (711 personas sentadas y de pie). Se estudiará una posible condición crítica de carga antes de la operación real para garantizar un comportamiento seguro tras la inauguración de la obra.

En específico se analizan dos terremotos: el del 22 de mayo de 1960 en Chile (9.5 Mw) correspondiente al más fuerte sismo registrado en el planeta, y el del 19 de septiembre de 1985 en México (8.1 Mw), significativo por la cercanía del epicentro con la construcción y su impacto en el país. Para este fin, se hace uso de técnicas de Ciencia de datos, como los algoritmos de agrupamiento y los de aprendizaje automático.

ESTADO DEL ARTE

Thai [2] realiza una revisión exhaustiva de las aplicaciones del Machine Learning en la ingeniería estructural, en el monitoreo de la salud estructural y en el diseño de estructuras de concreto. Además, ese autor presenta hallazgos, desafíos, recomendaciones y proporciona bases de datos y códigos de varios algoritmos como referencia.

Mangalathu S. y Jeon [3] utilizaron métodos de Machine Learning para predecir el modo de falla en columnas de puentes de concreto armado circular, con 311 muestras experimentales y tres tipos de fallas consideradas. Los métodos evaluados incluyeron análisis discriminante cuadrático, regla del vecino más cercano, árboles de decisión, bosques aleatorios, Naive Bayes y redes neuronales artificiales. Los resultados mostraron que las redes neuronales artificiales lograron una precisión del 91 % en la detección de fallas, superando a los otros métodos.

Nguyen et al. [4] presentaron un método para detectar daños en puentes usando cambios en la curvatura modal y Convolutional Neural Networks (CNN). El Gapped Smoothing Method (GSM) identifica la ubicación de los daños sin datos de referencia de estructuras sanas. Utilizando el puente Bonghi como ejemplo, se entrenó una CNN con imágenes del índice de daños del GSM para clasificar la ubicación del daño en una viga numérica. Los resultados muestran que la combinación de GSM y CNN puede detectar y localizar daños en puentes.

Mahmoudi et al. [5] realizan la detección de daños en edificios de muros de corte de concreto durante movimientos sísmicos. Utilizaron el algoritmo *K*-Nearest Neighbor ajustado con optimización bayesiana como modelo de predicción. Evaluaron el marco en tres edificios de referencia con datos generalizados, mostrando su eficacia en la identificación precisa de daños.

En el estudio de Abdelijaber et. al. [6] localizan daños estructurales en tiempo real utilizando redes neuronales convolucionales unidimensionales. Este enfoque destaca por su capacidad para extraer automáticamente características óptimas de las señales de aceleración y su eficiencia computacional como se demostró en experimentos realizados en un simulador de graderío.

En el estudio de Flah et al. [7] los algoritmos de Aprendizaje Automático (ML) en el Monitoreo de la Salud Estructural (SHM) son cada vez más relevantes por su capacidad para detectar daños en estructuras civiles. Con el crecimiento del Internet de las Cosas y Big data, estas aplicaciones se expandirán rápidamente. Esta revisión clasifica los algoritmos de ML en SHM basado en vibraciones e imágenes, analiza su eficacia y ofrece recomendaciones, destacando las brechas de conocimiento y las necesidades futuras de investigación.

Sarothi et al. [8] exploraron los modos de falla en conexiones atornilladas de doble corte en acero

estructural usando varios algoritmos en una base de datos de 455 resultados experimentales. Algoritmos como Random Forest, CatBoost, XGBoost y Gradient Boosting demostraron una precisión del 90-92 % en la clasificación de modos de falla, destacando la importancia de la relación entre la distancia al borde y el diámetro del perno en el rendimiento de estas conexiones.

Abdellatif et al. [9] identificaron grietas en imágenes de ingeniería civil mediante el algoritmo CrackIT, que ofrece enfoques basados en bloques y en píxeles. Proponen una nueva metodología que combina ambos para una detección precisa y tolerante al ruido. Evaluaron este enfoque en tres conjuntos de datos de imágenes de grietas, comparándolo con CrackIT y tres métodos de aprendizaje profundo. Su propuesta superó a los métodos existentes, reduciendo errores de discretización y mejorando la localización de grietas.

En línea con las investigaciones descritas, se realizó un estudio de dos fases para analizar la respuesta del tren interurbano México-Toluca frente a terremotos de gran magnitud y cargas excesivas, con el objetivo de evaluar su confiabilidad y seguridad para los futuros usuarios.

MÉTODOS DE APRENDIZAJE AUTOMÁTICO

Tal como se mencionó anteriormente, el estudio presentado en este artículo es en dos sentidos:

1. Descriptivo: Se analiza la estructura de los datos sintéticos previamente clasificados aplicando algoritmos de agrupamiento, para conocer la calidad de la distribución de datos: K-medias y EM
2. Predictivo: Con el agrupamiento obtenido se entrenaron los modelos de ML: Perceptrón Multicapa, Support Vector Machine y Árbol de Decisión para analizar su predicción.

Algoritmo K-medias

K-medias es un algoritmo de agrupamiento para lo que se requiere la especificación de una semilla k que determina la cantidad de grupos a obtener. El proceso general es el siguiente [10]:

1. Seleccionar los k centros iniciales.
2. Asignar las instancias (x) del conjunto de datos a su centroide más cercano (c) utilizando una métrica de similitud, que comúnmente es la distancia euclídea, Ec. (1).

$$d_{x,c} = \sqrt{\sum (x - c)^2} \quad (1)$$

3. Recalcular los centros considerando las instancias asignadas a ellos, Ec. (2).

$$\overline{c_j} = \frac{1}{c_j} \sum_{i=1}^{c_j} x_{ij} \quad (2)$$

4. Repetir los pasos 2 y 3 hasta que los centroides convergen.

Algoritmo EM

Expectation-Maximitation (EM) es un algoritmo no supervisado que identifica las distribuciones gaussianas de los datos mediante la generación de grupos considerando los siguientes pasos [11] :

1. Inicialización aleatoria del número de grupos esperados, así como los parámetros de las distribuciones gaussianas, considerando el rango de valores posibles de los datos.
2. Expectation. Realizar el cálculo de las distancias de los datos con respecto a las medias de las Gaussianas para determinar la distribución de probabilidad de los datos.
3. Maximitation. La distribución de probabilidad obtenida en la etapa anterior, son modificados mediante el cálculo de las nuevas medias y matrices de covarianza.
4. Repetir los pasos 1 y 2 hasta que no haya cambios en la asignación de datos a los grupos.

Perceptrón Multicapa

El Perceptrón multicapa (PM) es un modelo de clasificación supervisada popular debido a su robustez. Su topología consiste en tres tipos de capas: de entrada, oculta y de salida (ver Figura 1).

El algoritmo más utilizado para entrenar el PM es el Backpropagation, que incluye dos etapas principales: propagación hacia adelante y propagación hacia atrás de la señal [12]. El proceso general es el siguiente:

1. Inicialización de los pesos de las conexiones y los parámetros libres de la red (la razón de aprendizaje, factor momento, mínimo error esperado, entre otros).
2. Propagación hacia adelante: Cada instancia del conjunto de datos se presenta a la red y se propaga la señal hasta la capa de salida, donde se determina el error respecto a la salida esperada.

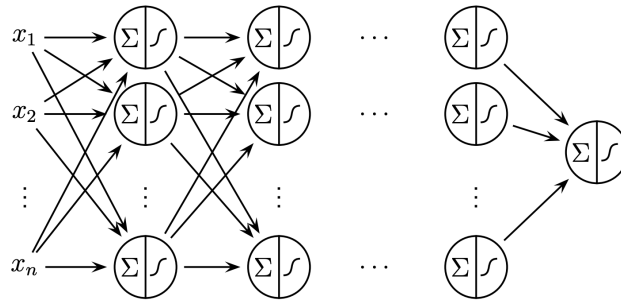


FIGURA 1. Topología típica de un PM. $[x_1 \dots x_n]$ representan el vector de entrada, Σ Representa la regla de propagación. f Denota la función de activación.

3. Propagación hacia atrás: La señal se propaga de la capa de salida y hasta la capa de entrada realizando el ajuste de los pesos considerando el gradiente descendente para minimizar el error cuadrático medio observado con cada instancia.
4. Los pasos 2 y 3 son repetidos hasta la convergencia de la red a un mínimo error obtenido.

Support Vector Machine

El objetivo del *Support Vector Machine* (SVM) es identificar el hiperplano que separa de forma óptima las clases de un problema, que tenga en consideración maximizar la distancia existente entre éste y las instancias más cercanas a él (ver Figura 2).

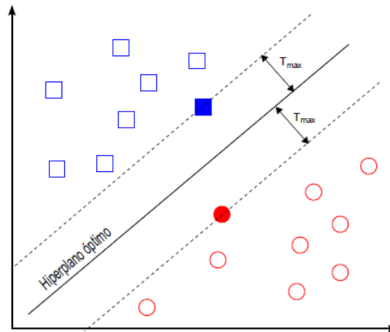


FIGURA 2. Hiperplano óptimo en un problema de dos clases.

El proceso que sigue el SVM consta de los siguientes pasos [13]:

1. Realizar el mapeo de los datos. En caso de ser necesario, transformar los datos de entrada a un espacio de dimensión alta (truco de kernel).
2. Identificar el hiperplano que mejor separa las clases en el problema.

3. Encontrar los vectores de soporte, que corresponden a dos hiperplanos paralelos al hiperplano que separa las clases, maximizando la distancia a la instancia más cercana a él.

Árbol de Decisión

El árbol de decisión es un clasificador supervisado que realiza particiones repetidas en el conjunto de datos para agrupar elementos de una sola clase. Su construcción sigue cuatro pasos.

1. Para cada atributo calcular la entropía y la ganancia de información normalizada.
2. Tomar como padre al atributo que tenga la ganancia más alta y crear un nodo de decisión que lo divida considerando los valores que el atributo adquiere.
3. Repetir en las sablistas obtenidas por división y agregar estos nodos como hijos.
4. Finalizar cuando se llegue a las hojas.

Se utilizó el lenguaje de programación Python con ayuda de bibliotecas fundamentales para el análisis de datos, la manipulación y la visualización, como lo son pandas, numpy y matplotlib.

METODOLOGÍA

Para dar solución al problema planteado se siguió proceso presentado en la Figura 3.

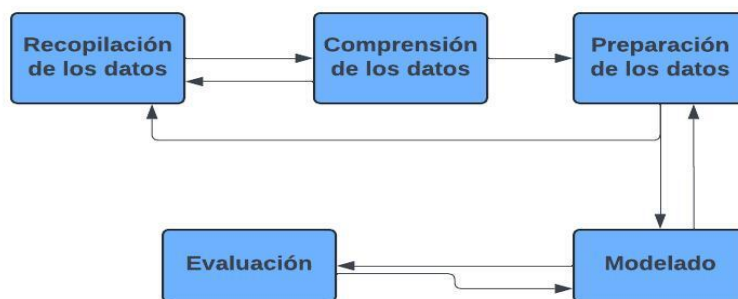


FIGURA 3. Estrategia de solución utilizada.

Compresión de los datos

Utilizando como valores de referencia los relacionados a la estructura física del tren interurbano México-Toluca [14], en un estudio previo [15] se diseñó un modelo estructural utilizando el programa Structural Analysis Program (SAP 2000). A partir de este modelo se generaron distintas simulaciones considerando

los sismos de Chile y de México respectivamente con el fin de observar la variabilidad y el comportamiento del conjunto de variables. Esto dio paso a la construcción de una Red Bayesiana con 14 nodos, que relaciona a las variables de sismo y carga viva máxima, con los elementos mecánicos de una pila de soporte y sus desplazamientos lineales y angulares. Después se ajustó la red bayesiana propuesta inicialmente modificando la variable relacionada con los sismos para considerar movimientos telúricos más intensos, para someter la estructura a un mayor esfuerzo y determinar la probabilidad de falla.

Para ambos sismos, se consideraron las características bajo el criterio de las variables del sismo, la carga viva máxima y los elementos mecánicos de una pila de apoyo que es donde normalmente ocurren las fallas de un viaducto. En el estudio, las 7 características del conjunto de datos son: factor de amplificación del sismo (GF), peso total de los pasajeros más peso de los ejes del tren (Wt), máximo esfuerzo por carga axial (P), máximo esfuerzo cortante en dirección X y Y (V_x , V_y), máximo momento de flexión respecto a X y Y (M_{x1} , M_{y1}), máximo momento de torsión (T), desplazamientos lineales en U_x , U_y , U_z , y giros con respecto a los tres ejes R_x , R_y , R_z (Figura 4).

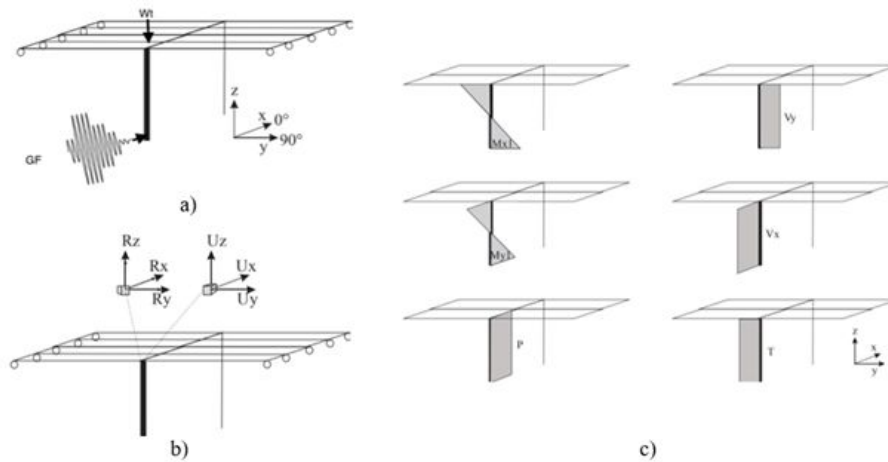


FIGURA 4. Variables utilizadas en la pila de un apoyo del viaducto elevado del tren México-Toluca. a) Variables del sismo, peso total del tren y de los pasajeros, b) Variables para los desplazamientos en el nodo superior y c) Variables utilizadas para los esfuerzos.

Preparación de los datos

En esta etapa se buscó mejorar la calidad del conjunto de datos, para ello, con la función *duplicated* en Python, se identificó que el 17.55 % de los registros del conjunto de datos del sismo de Chile eran duplicados, por lo que se eliminaron. Para detectar los outliers se utilizaron gráficas de caja para describir visualmente la distribución de datos [16] y posteriormente se reemplazaron los valores atípicos con los obtenidos del cálculo de los máximos o mínimos del rango intercuartil del conjunto de datos (diferencia entre el tercer y el primer cuartil de la columna analizada). Como resultado, se obtuvo la distribución

mostrada en la Tabla 1. Cada conjunto de datos está compuesto por 4 clases correspondientes a los factores de amplificación del sismo.

Tabla 1. Característica de los conjuntos de datos utilizados.

Conjunto de datos	Atributos	Clase 1	Clase 2	Clase 3	Clase 4
Sismo Chile	32	173	173	272	202
Sismo México	16	2314	2674	2557	2454

Modelado

Los modelos construidos se basaron en varios de los algoritmos más utilizados en el aprendizaje automático utilizando el software Weka en su versión 3.8.6, considerando los valores por default:

1. EM: maxIterations: 500, minStdDev: 1.0E-6, numClusters: 4, numExecutionSlots: 1, numFolds:10, numKMeansRuns: 10, seed: 100.
2. K-means: canopyMaxNumCanopiesToHoldInMemory:100, numClusters: 4, numExecutionSlots: 1, maxIterations:500, seed:10, canopyMinimumCanopyDensity: 2.0.
3. Multilayer Perceptron: learningRate: 0.3, momentum: 0.2, numDecimalPlaces:2, seed: 0, training Time: 500, validationSetSize: 0, validationThreshold: 20.
4. Arbol J48: confidenceFactor: 0.25, minNumObj: 2, numDecimalPlaces: 2, numFolds: 3, seed: 1.
5. SVM: c: 1, epsilon: 1.0E-12, numDecimalPlaces: 2, numFolds: -1, randomSeed: 1, toleranceParameter: 0.001.

Evaluación

En los experimentos se usó validación cruzada con 10 particiones, empleando el 80 % de los datos para entrenamiento y el 20 % para validación. Se evaluó el rendimiento de los modelos mediante dos medidas comunes en ciencia de datos: precisión general y media geométrica, derivadas de una matriz de confusión como la de la Tabla 2.

Tabla 2. Muestra una matriz de 2x2 para un problema de dos clases.

	Predicción Clase (+)	Predicción Clase (-)
Clase real (+)	Verdadero positivo (VP)	Falso negativo (FN)
Clase real (-)	Falso positivo (FP)	Verdadero negativo (VN)

En Tabla 2, VP y VN representan el número de instancias clasificadas por el modelo, mientras que FP y FN corresponden a errores cometidos de ambas clases. De este modo, la precisión general se expresa como Ec. (3):

$$Pg = \frac{VP + VN}{VP + FN + VN + FP} \quad (3)$$

Debido al desequilibrio existente en los datos, la utilización de la media geométrica Ec. (4), es ampliamente utilizada:

$$g = \sqrt{a^+ * b^-} \quad (4)$$

Donde: a^+ es la precisión de la clase positiva = $\frac{VP}{VP+FN}$ y a^- es la precisión de la clase negativa = $\frac{VN}{VN+FP}$

EXPERIMENTACIÓN

En el objetivo de esta investigación, se persiguen dos resultados: validar la distribución de instancias en la generación inicial de los conjuntos de datos y estimar el comportamiento de la estructura ferroviaria del tren interurbano si se presenta algún sismo de magnitud similar al del caso de estudio.

Agrupamiento

Para validar la distribución de datos en el conjunto creado con simulaciones en SAP 2000, se aplicaron los algoritmos EM y K -medias. La Tabla 3 muestra la distribución de instancias tras el agrupamiento del conjunto de datos de Chile. Las filas representan los valores de referencia y las columnas la asignación del agrupamiento.

Del agrupamiento obtenido se pueden hacer varios comentarios. La integración de la clase 3 en ambos agrupamientos es consistente y libre de ruido, como se observa en la intersección de la clase 3 con el grupo 3, donde los valores coinciden con los de referencia (Tabla 1). Para K -medias, la clase 2 prácticamente desaparece, asignándose a la clase 1. Esto se evidencia en las intersecciones entre las clases 2 y el grupo 2, así como entre la clase 1 y el grupo 1. Al momento de analizar el agrupamiento

obtenido con el conjunto de datos de sismo en México, no es posible encontrar una correspondencia lógica entre los grupos identificados y las clases existentes (Tabla 4). En todos los grupos obtenidos se encuentran instancias de las cuatro clases, lo que indica un problema en la calidad de los datos fuente.

Tabla 3. Distribución del agrupamiento del sismo en Chile.

	Grupo 1		Grupo 2		Grupo 3		Grupo 4	
	EM	k-medias	EM	k-medias	EM	k-medias	EM	k-medias
Clase 1	105	173	70	0	0	0	0	0
Clase 2	68	173	103	0	0	0	0	0
Clase 3	0	0	0	0	272	272	0	0
Clase 4	0	0	0	89	0	0	202	113
Total	173	347	173	89	272	272	202	113

Tabla 4. Distribución del agrupamiento del sismo en México.

	Grupo 1		Grupo 2		Grupo 3		Grupo 4	
	EM	k-medias	EM	k-medias	EM	k-medias	EM	k-medias
Clase 1	1241	126	1219	227	996	274	608	358
Clase 2	297	1863	562	1789	636	1343	841	758
Clase 3	639	188	580	345	447	462	283	616
Clase 4	137	137	313	313	478	478	722	722
Total	2314		2674		2557		2454	

Clasificación

La Tabla 5 presenta el promedio de precisión general y la media geométrica, destacando en negrita los mejores resultados. Utilizar las distribuciones obtenidas con los algoritmos de agrupamiento mejora sustancialmente la clasificación, con poca variación entre algoritmos de clasificación. En su mayoría, los mejores resultados son obtenidos con el algoritmo EM.

Tabla 5. Resultados de la clasificación.

CD	Medida	PM		SVM		J48	
		Chile	México	Chile	México	Chile	México
Original	Pg	80.8	40.68	76.9	40.74	80.2	37.33
	g	85.1	57.2	82	57.7	86.6	54.6
EM	Pg	97.8	99.99	97.6	99.99	98.3	99.99
	g	98.3	99.9	98.1	99.9	98.7	99.9
k-medias	Pg	99.3	99.5	99	99.62	98.7	99.12
	g	99.1	99.5	98.8	99.6	98.5	99.1

La Tabla 6 muestran en promedio la matriz de confusión obtenida al utilizar los modelos de ML con el conjunto de datos original, es decir, sin agrupar.

Tabla 6. Distribución del agrupamiento del sismo en México.

Clase	Sismo Chile				Sismo México			
	Perceptrón Multicapa		Support Vector Machine		Clasificador J48			
	Chile	México	Chile	México	Chile	México	Chile	México
Perceptrón Multicapa								
Clase 1	242	22.33	0	0	2027.7	299.3	72.7	54.7
Clase 2	30	162.3	0	2	203.7	3098.7	128.7	157.7
Clase 3	0	0.33	194.7	0.7	109.3	357.3	1319	253.3
Clase 4	0	4	1.3	161.3	31	206.7	120	1560.3
Support Vector Machine								
Clase 1	232.7	31.7	0	0	2159.7	214.7	46.3	33.7
Clase 2	31.7	159	0	3.7	351.7	3006.7	96	134.3
Clase 3	0	0	194.7	1	210.7	297.7	1295.7	235
Clase 4	0	3	1.7	162	72.7	180.7	115	1549.7
Clasificador J48								
Clase 1	256	8.3	0	0	2021	225.7	145.3	62.3
Clase 2	46	145.7	0.3	2.33	219.3	2962.7	239.3	167.3
Clase 3	0	0	194	1.67	148.3	250.7	1399.3	240.7

Tabla 6. Distribución del agrupamiento del sismo en México (...continuada).

Clase	Sismo Chile				Sismo México			
	Perceptrón Multicapa		Support Vector Machine		Clasificador J48			
	Chile	México	Chile	México	Chile	México	Chile	México
Clase 4	0.3	2	1.7	162.7	54	152	213.7	1498.3

Las matrices de confusión anteriores destacan el traslape de clases en el conjunto de datos del sismo de México, lo que justifica la baja precisión observada en los tres clasificadores utilizados.

Una estrategia estadística común para analizar la concordancia en los resultados es el análisis del coeficiente Kappa [17], que compara la concordancia observada en un conjunto de datos, respecto a la que podría ocurrir por mero azar. Estos resultados se muestran en la Tabla 7.

Tabla 7. Coeficiente Kappa obtenido con los diferentes modelos utilizados.

CD	Sismo	PM	SVM	J48
Original	Chile	0.7401	0.6874	0.7317
	México	0.2066	0.2115	0.1637
EM	Chile	0.9703	0.9669	0.9768
	México	0.9999	0.9999	0.9999
K-medias	Chile	0.9893	0.9857	0.9803
	México	0.9918	0.9937	0.9855

Los valores de kappa varían de -1 a $+1$, mientras más alto sea el valor, más fuerte es la concordancia. De la Tabla 7, es posible ver que los resultados obtenidos con el conjunto de datos del sismo de Chile cuentan con un grado de acuerdo, reforzando las conclusiones obtenidas.

CONCLUSIONES

Los resultados muestran que el análisis de clúster revela un severo traslape de clases en los datos sintéticos originales del sismo en México comparado con Chile. Esto se debe a la dificultad para encontrar una correspondencia lógica entre los grupos identificados y las clases en los datos de México. La distribución obtenida por los algoritmos de agrupamiento cuestiona la distribución generada inicialmente con el software SAP 2000 y se espera que afecte el rendimiento de los modelos al intentar aprender a clasificar.

Por otro lado, en lo relacionado a la clasificación considerando tanto las clases inicialmente definidas, así como las identificadas con los algoritmos de agrupamiento, fue posible observar que, al ser entrenados con los conjuntos derivados del agrupamiento, mejora su desempeño, principalmente con el algoritmo EM.

Las líneas de estudio abiertas se dirigen hacia dos aspectos: aplicar técnicas más profundas de limpieza de datos para resolver el problema de traslape de clases antes de la clasificación y utilizar métodos para generar datos sintéticos que conserven las distribuciones estadísticas, aumentando el volumen de datos y mejorando el rendimiento de los clasificadores.

Es importante tener en cuenta que las circunstancias sobre las que se hizo el estudio de la estructura fueron muy exigentes, y las probabilidades de que se produzcan las cargas que se analizaron son escasas. La estructura está dentro de los parámetros de seguridad establecidos, aunque se requiere más investigaciones en futuros trabajos para identificar las condiciones críticas de seguridad.

Finalmente, es necesario comentar que este artículo forma parte de un proyecto de investigación (UAEMex 6523/2022 CIA) que pretende establecer la probabilidad de falla de un “tramo tipo” del viaducto elevado del tren México-Toluca, sujeto simultáneamente a sismo y carga viva máxima. Publicaciones posteriores complementarán, entonces, la hasta aquí presentada.

REFERENCIAS

- [1] Secretaría de Obras y Servicios de la CDMX. *Tren El Insurgente*. Recuperado el 15 de junio de 2024, de <https://www.obras.cdmx.gob.mx/proyectos/tren-interurbano-cdmx>
- [2] H.-T. Thai. Machine learning for structural engineering: A state-of-the-art review. *Structures*, vol. 38, pp. 448–491, Apr. 2022. DOI: 10.1016/j.istruc.2022.02.003
- [3] S. Mangalathu and J.-S. Jeon. Machine Learning-Based Failure Mode Recognition of Circular Reinforced Concrete Bridge Columns: Comparative Study. *Journal of Structural Engineering*, vol. 145, pp. 04019104, Oct. 2019. DOI: 10.1061/(ASCE)ST.1943-541X.0002402
- [4] D. H. Nguyen, Q. B. Nguyen, T. Bui-Tien, G. De Roeck, and M. Abdel Wahab. Damage detection in girder bridges using modal curvatures gapped smoothing method and Convolutional Neural Network: Application to Bo Nghi bridge. *Theoretical and Applied Fracture Mechanics*, vol. 109, pp. 102728, Oct. 2020. DOI: 10.1016/j.tafmec.2020.102728

- [5] H. Mahmoudi, M. Bitaraf, M. Salkhordeh, and S. Soroushian. A rapid machine learning-based damage detection algorithm for identifying the extent of damage in concrete shear-wall buildings. *Structures*, vol. 47, pp. 482–499, Nov. 2022. DOI: 10.1016/j.istruc.2022.11.041
- [6] O. Abdeljaber, O. Avci, S. Kiranyaz, M. Gabbouj, and D. J. Inman. Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks. *Journal of Sound and Vibration*, vol. 388, pp. 154–170, Feb. 2017. DOI: 10.1016/j.jsv.2016.10.043
- [7] M. Flah, I. Nunez, W. Ben Chaabene, and M. L. Nehdi. Machine Learning Algorithms in Civil Structural Health Monitoring: A Systematic Review. *Archives of Computational Methods in Engineering*, vol. 28, no. 4, pp. 2621–2643, Jun. 2021. DOI: 10.1007/s11831-020-09471-9
- [8] S. Sarothi, K. Ahmed, N. Imtiaz, A. Ahmed, and M. Nehdi. Machine Learning-Based Failure Mode Identification of Double Shear Bolted Connections in Structural Steel. *Engineering Failure Analysis*, vol. 139, pp. 106471, 2022. DOI: 10.1016/j.engfailanal.2022.106471
- [9] M. Abdellatif, H. Peel, A. G. Cohn, and R. Fuentes. Combining block-based and pixel-based approaches to improve crack detection and localization. *Automation in Construction*, vol. 122, pp. 103492, Feb. 2021. DOI: 10.1016/j.autcon.2020.103492
- [10] A. H. Alkadhwi and O. Adelaja. Data mining application using clustering techniques (K-Means algorithm) in the analysis of student's result. *Journal of Multidisciplinary Engineering Science Studies (JMESS)*, vol. 5, no. 5, pp. 2458–925, 2019. ISSN: 2458-925X.
- [11] B. Pani, J. Klemenc, and M. Nagode. Improved Initialization of the EM Algorithm for Mixture Model Parameter Estimation. *Mathematics*, vol. 8, no. 3, p. 373, 2020. DOI: 10.3390/math8030373
- [12] D. A. Alboaneen, H. Tianfield, and Y. Zhang. Sentiment analysis via multi-layer perceptron trained by meta-heuristic optimisation. In *2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 4630–4635. DOI: 10.1109/BigData.2017.8258507
- [13] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020. DOI: 10.1016/j.neucom.2019.10.118
- [14] J. Jiménez. Función del enlace técnico dentro de la obra del tren Interurbano México Toluca. *Memoria de Experiencia Laboral*, Ingeniería Civil, Facultad de Ingeniería, Universidad Autónoma del Estado de México, 2017.

- [15] D. J. Delgado Hernández, R. A. Mejía Nava, D. De León Escobedo, J. C. Arteaga Arcos, M. Á. Mendoza Lugo, and J. F. García Zamora. Análisis de riesgos del tren México-Toluca sujeto simultáneamente a sismo y carga viva máxima. Informe del proyecto 6523/2022CIA. SIEA, Universidad Autónoma del Estado de México, 2023.
- [16] R. Arora and S. Suman. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. *International Journal of Computer Applications*, vol. 54, pp. 21–25, 2012.
- [17] J. R. Dettori and D. C. Norvell. Kappa and Beyond: Is There Agreement? *Global Spine Journal*, vol. 10, no. 4, pp. 499–501, 2020. DOI: 10.1177/2192568220911648
- [18] K. Stapor, P. Ksieniewicz, S. García, and M. Woniak. How to design the fair experimental classifier evaluation. *Applied Soft Computing*, vol. 104, p. 107219, Jun. 2021. DOI: 10.1016/j.asoc.2021.107219
- [19] J. B. Rollins. Metodología Fundamental para la Ciencia de Datos. *IBM, Armonk, NY, Technology, IMWI4828-COES-01*, 2015.

Un modelo de Red Bayesiana para datos cualitativos de Vaginosis Bacteriana en mujeres embarazadas

A Red Bayesian Model for Qualitative Data of Bacterial Vaginosis in Pregnant

Bautista Hernández, C. M.^{1*} , Canul Reich, J.¹ , López Ramírez, C.¹, De la Cruz Hernández, E.¹
(*carlos.bta@icloud.com)

Recibido: 13 de junio de 2024 Aceptado: 28 de junio de 2024

RESUMEN

La Vaginosis Bacteriana BV es un factor de riesgo para el parto prematuro. La Organización Mundial de la Salud OMS, destaca que a pesar de que la BV ya se considera una afección más común y de no tratarse puede provocar problemas en el embarazo como aborto espontáneo y parto prematuro así como mayor riesgo de otras infecciones de transmisión sexual. Las técnicas computacionales que parten de la Inteligencia Artificial como Redes Bayesianas Causales CBNs, se han utilizado en diferentes estudios con el objetivo de análisis de diagnóstico de enfermedades. Este estudio tiene la finalidad de compartir cómo una CBN puede identificar que bacterias coexisten y la relación que tienen frente a un diagnóstico

positivo DxBV+. Los datos que se utilizaron para este estudio fueron una recolección de datos reales por parte de personal investigador de la División Académica Multidisciplinaria de Comalcalco DAMC formando un *dataset* de casos positivos, negativos e indeterminados de BV en mujeres embarazadas. El resultado que se obtuvo dio un modelo gráfico causal que es presentado como una CBN y su relación de probabilidad condicional que detalla el porcentaje de relación con el DxBV+.

Palabras clave: *Vaginosis Bacteriana BV, Mujeres Embarazadas, Infección Genital, Redes Bayesianas BN.*

¹Universidad Juárez Autónoma de Tabasco. Carr. Cunduacan a Jalpa de Méndez, KM 1, C.P. 86690.



ABSTRACT

BV Bacterial Vaginosis is a risk factor for preterm birth. The World Health Organization WHO, highlights that although BV is already considered a more common condition and if left untreated it can cause problems in pregnancy such as spontaneous abortion and premature birth as well as an increased risk of other sexually transmitted infections. Artificial Intelligence computational techniques such as Causal Bayesian Networks CBNs have been used in different studies with the objective of disease diagnostic analysis. This study aims to share how a CBN can identify which bacteria coexist and the relationship they have with a positive DxBV+ diagnostic. The data that was used for this study is based on a collection of real data by research personnel from the Multidisciplinary Academic Division of Comalcalco DAMC, forming a dataset of positive, negative and indeterminate cases of BV in pregnant women. The result obtained gave a causal graphic model that is presented as a CBN and its conditional probability table that details the percentage of relationship with DxBV+.

Keywords: *BV Bacterial Vaginosis, Genital Infection, Pregnant Women, Bayesian networks.*

INTRODUCCIÓN

La Vaginosis Bacteriana (BV, por sus siglas en inglés) es una infección genital femenina de las más comunes y el impacto sobre la salud de una mujer es significativo. La BV es una enfermedad genital que afecta tanto a las mujeres sanas y embarazadas [1]. Se sabe que hay diferentes condiciones genéticas y otras condiciones de ambiente en las que puede iniciar la infección. Un elevado pH vaginal facilita el crecimiento de organismos de transmisión sexual [2].

Se tienen indicios que enfermedades como la clamidia, el Virus del Papiloma Humano VPH, el Síndrome de Inmunodeficiencia Humana VIH y otras de transmisión sexual pueden estar en ocasiones asociadas a la BV [3]. La definición en la que coinciden en diversos artículos es que la infección que comienza a dar lugar a la BV se origina debido a un disturbio en la flora vaginal normal y la interacción de diversos agentes patógenos [4].

La BV se conoce que es la causa más común de flujo vaginal en mujeres en edad reproductiva y está asociada a una gran cantidad de condiciones incluyendo infección postquirúrgica, enfermedad pélvica inflamatoria, parto pretérmino y enfermedades como el VPH y VIH [5]. Debido a los problemas que ocasiona la BV se plantea experimentar cómo una Red Bayesiana Causal puede dar resultados sobre qué bacterias coexisten en el diagnóstico positivo de la BV.

La Red Bayesiana Causal es un algoritmo usado en *Machine Learning* (ML) con la cual pueden crearse modelos gráficos. El uso de los modelos gráficos ayuda a realizar modelos predictivos [6] mediante un

gran procesamiento de datos [7].

Los datos que se usan en este experimento fueron recolectados de pacientes por parte de personal de investigación de la División Académica Multidisciplinaria de Comalcalco (DAMC) de la Universidad Juárez Autónoma de Tabasco UJAT [8]. El *dataset* contiene tres clases: positivos, negativos e indeterminados para el diagnóstico positivo de la BV en mujeres embarazadas.

Se realizaron los experimentos en búsqueda de causas detonantes que originan la Vaginosi s Bacteriana Positiva BV+ en el *dataset* mencionado anteriormente. Como resultado se obtuvo un Modelo Gráfico Causal que fue creado con la intervención del algoritmo *Expectation – Maximization* (EM) para aprender la estructura de la Red Bayesiana y el algoritmo para Redes Bayesianas Causales conocido como *Fast Causal Inference* (FCI) para optimizar las probabilidades.

MATERIALES Y MÉTODOS

Redes Bayesianas

La creación de las Redes Bayesianas (BN, por sus siglas en inglés) tienen como fundamento el teorema de bayes para identificar el grado de relaciones entre las clases. El Teorema de Bayes tiene sus principios en los años 1700 a través de la transformación de diversas teorías de probabilidad [9]. Fue así que Tomas Bayes (1701-1761) dio origen a este teorema basándose en las investigaciones anteriores. Las BN se representan con un Gráfico Acíclico Dirigido (DAG, por sus siglas en inglés).

Teorema de Bayes

El Teorema de Bayes tiene sus principios en el Teorema de la Probabilidad Total: la probabilidad que ocurran los sucesos A y B , es la probabilidad de que ocurra A multiplicada por la probabilidad de que ocurra B , que está condicionada por la ocurrencia de A , esto es la fórmula de la Ec. (1):

$$P(A \cap B) = P(A) \cdot P(B|A) \quad (1)$$

La fórmula mostrada en la Ec. (1) es la base para la fórmula mostrada en la Ec. (2) [9]:

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)} \quad (2)$$

Donde:

$P(A|B)$ es la probabilidad de que ocurra un evento A dado que ha ocurrido un evento B , también conocida como la probabilidad posterior.

$P(A)$ es la probabilidad de que ocurra un evento A sin tener en cuenta ninguna información sobre B conocida también como probabilidad previa o a priori.

$P(B|A)$ es la probabilidad de que ocurra el evento B , dado que ha ocurrido el evento A , conocida como probabilidad condicional.

$P(B)$ es la probabilidad de que ocurra el evento B independientemente de cualquier otra información. También conocida como probabilidad marginal [9].

Interpretación y aplicaciones del Teorema de Bayes

El Teorema de Bayes describe la relación entre la probabilidad condicional de dos eventos y permite actualizar la probabilidad de una hipótesis basándose en nueva evidencia. La fórmula de Bayes $P(A|B)$ permite calcular la probabilidad de A dado que B ha ocurrido, o de manera inversa la probabilidad de B dado que ha ocurrido el evento A , $P(B|A)$. El Teorema de Bayes se utiliza ampliamente en medicina y otras disciplinas [9].

Inferencia Causal

La Inferencia Causal (CI, por sus siglas en inglés) puede parecer simple, pero es una tarea científica compleja ya que se basa en la triangulación de evidencias de múltiples fuentes y en la aplicación de una variedad de enfoques metodológicos. La CI ha sido atribuida a cómo poder predecir resultados mediante eventos causales. Judea Pearl menciona que a través de la estadística se han proporcionado herramientas de predicción de probabilidades [10].

El objetivo de la CI es inferir no solo creencias o probabilidades en condiciones estáticas, sino también la dinámica de las creencias en condiciones cambiantes, cambios inducidos por tratamientos, nuevas políticas u otras intervenciones externas [11]. La CI recaba conocimiento que toma para la predicción de sucesos mediante el almacenamiento y procesamiento en una memoria.

Modelos Gráficos Causales

El estudio obtuvo como resultado la creación de un Modelo Gráfico Causal (MGC), mediante la combinación de BN y CI. Las dos técnicas aplicadas dieron un resultado de Redes Bayesianas Causales (CBN, por sus siglas en inglés) [3].

Las CBN se representan mediante un MGC el cual es una representación visual y matemática de las relaciones causales que se pueden encontrar entre un conjunto de variables. La creación de un MGC requiere grafos con inferencia causal, además cuenta con nodos que representan variables por ejemplo A, B, C y aristas que representan las relaciones causales [11], ver Ec. (3):

$$A \rightarrow B \rightarrow C \quad (3)$$

Donde: A causa B y B causa C

Redes Bayesianas Causales

Las CBNs se pueden interpretar con DAGs, donde los nodos representan variables y los arcos relaciones causales. En un MGC las CBNs representan relaciones causas-efecto, por lo que implican supuestos más fuertes que tener simplemente una BN.

La estructura de una CBN se puede construir de dos maneras. La primera manera, es aplicando algoritmos para las BNs y realizar el DAG de manera manual con base en estas relaciones; la segunda, es aplicando un algoritmo que realice las relaciones entre variables con base en los cálculos de probabilidad. A las clases objetivo se les debe fijar el valor de una o más variables y establecer su efecto en las demás, esto es el razonamiento causal [3].

Algoritmo Expectation-Maximization (EM)

EM es una técnica iterativa para estimar los parámetros de los modelos probabilísticos que permite generar una estructura de BN aplicándolo en herramientas de desarrollo de software [12]. El algoritmo EM se utiliza para crear un modelo que relaciona los datos observados y ausentes con elecciones de decisores. El modelo predice las elecciones producidas de los datos observados [13].

La idea básica del algoritmo EM es formar la estructura con los mejores parámetros de un modelo, cuyos pasos en este estudio son:

- **Inicialización:** Se inicia con suposiciones aleatorias o neutrales sobre las probabilidades.
- **Expectation (E):** Se utilizan las probabilidades condicionales de las variables con las que se cuenta para calcular las probabilidades marginales de las variables que parecen no tener una intervención. Por ejemplo, si se conoce la probabilidad de que llueva dado el clima y la temperatura, se procede a calcular la probabilidad de que llueva sin tener en cuenta estos factores.
- **Maximization (M):** Se utilizan las estimaciones de las probabilidades marginales para actualizar las probabilidades condicionales de las variables. Por ejemplo, si se observa que la lluvia es más probable cuando hace frío y está nublado, se ajustan las probabilidades condicionales de lluvia dada la temperatura y el clima para reflejar esto.
- **Repetir:** Se repiten los pasos E y M hasta que las probabilidades condicionales converjan a valores estables. Para mayor detalle ver [11].

Dataset

En el *dataset* se utiliza el algoritmo EM para generar la estructura con base en las probabilidades condicionales de todas las variables cualitativas. El conjunto de datos que se utilizó para este estudio proviene de la recolección de datos del investigador científico Dr. Erick de la Cruz Hernández perteneciente DAMC de la UJAT.

El *dataset* se encuentra constituido de 11,484 instancias con una distribución de 87 columnas y 132 filas. Se encuentran dos categorías diferentes: variables cualitativas que son 59 incluyendo dos variables clases u objetivo y variables cuantitativas que son 28. La variable clase se encuentra con el nombre de DxVB+ que es el diagnóstico positivo para la Vaginosis Bacteriana y se encuentra clasificada de la siguiente manera: para los casos positivos se representan con 1, para casos negativos con 2 y, finalmente, para casos indeterminados con 3.

Software

Se utilizó el lenguaje de programación R, el cual se ejecutó en el Entorno de Desarrollo Integrado (IDE, por sus siglas en inglés), *RStudio* en su versión 2023.12.0.369.

ESTUDIO EXPERIMENTAL

El método utilizado en la experimentación consta de las etapas a continuación descritas:

- Exploración de datos

En la variable clase se encuentran tres categorías descritas en la Tabla 1.

Tabla 1. Análisis de las instancias de la variable clase.

Casos	DxVB+
Positivos (1)	32 registros
Negativos (2)	5 registros
Indeterminados (3)	95 registros

- Preprocesamiento de los datos

En esta fase se realizaron los procedimientos para llevar a cabo la adecuación de las variables del *dataset*. Se utilizaron cuatro procedimientos principales que constan de: a) selección de datos, b) limpieza de los datos, c) transformación de los datos, d) imputación e integración de datos faltantes.

- a) Selección de los datos. El *dataset* que constaba de 87 variables y se redujo a 77 variables, quedo conformado por 28 variables cuantitativas y 47 variables cualitativas que fueron las utilizadas. Se descartaron variables como datos generales de los pacientes y otras clases de estudio como el VPH.
- b) Limpieza de datos: Se trataron datos con información anómala o con errores de captura.
- c) Transformación de datos: Algunas variables numéricas necesitaron ser transformadas a variables categóricas para poder trabajar con los datos cualitativos presentes en el *dataset*.
- d) Imputación e integración de los datos: Se tomaron las variables cualitativas y se realizó un procedimiento con base en la técnica estadística MODA [12], para rellenar los datos faltantes.

La MODA aplicada, se realizó con relación a la variable clase, es decir a la variable del diagnóstico positivo de la BV. En la Tabla 2 se muestra la lógica que se precisó en este método, en dos variables cualitativas con un ejemplo de 11 instancias por variable.

Tabla 2. Ejemplo de imputación, datos originales de *dataset* de BV.

Comezón genital	Ardor	Diagnóstico BV+
2	2	3
NA	NA	1
NA	NA	3
NA	NA	3
1	2	3
NA	NA	3
2	2	3
2	2	3
2	2	1
2	2	2
2	2	2

Con base en la Tabla 2, cuando el diagnóstico BV+ es 1, 2 o 3, la MODA para Comezón genital es 2 y para Ardor es 2, por lo que se procede a la imputación mostrados en la Tabla 3.

Tabla 3. Ejemplo de imputación, datos procesados.

Comezón genital	Ardor	Diagnóstico BV+
2	2	3
2	2	1
2	2	3
2	2	3
1	2	3
2	2	3
2	2	3
2	2	3
2	2	1
2	2	2
2	2	2

En la Tabla 3 se integran los datos al *dataset* y se concluye con el ejemplo de lo realizado en la etapa de preprocesamiento o preparación de los datos.

■ Experimentación con los datos

En la experimentación se tomó el resultado final del preprocesamiento de los datos descritos en la sección anterior. Se implementó el algoritmo EM con las 47 variables cualitativas del *dataset* para obtener la BN, el algoritmo EM se utilizó con los parámetros que tiene predefinidos por defecto. Para el análisis de resultados se ajustaron los parámetros para obtener como resultado solo las variables que se encuentran con un rango mayor al 50 % de relación de la probabilidad condicional que calcula el algoritmo EM.

El algoritmo EM hace los cálculos de probabilidad condicional entre las variables cualitativas del *dataset* y la variable clase identificada como diagnóstico positivo de la BV. Inicialmente se obtuvo una BN creada con la implementación del algoritmo EM en combinación con todas las variables cualitativas del *dataset*. El resultado inicial de esta BN experimentada se muestra en la Figura 1.

En la Figura 1 se pueden apreciar nodos y arcos dirigidos. La dirección del arco sugiere una interpretación causal o probabilística entre los nodos. Considerando a la variable *Lactobacillus inners* como el nodo (A) y a la variable Dolor como el nodo (B), se interpreta que el nodo A tiene un efecto causal o una influencia directa sobre el nodo B.

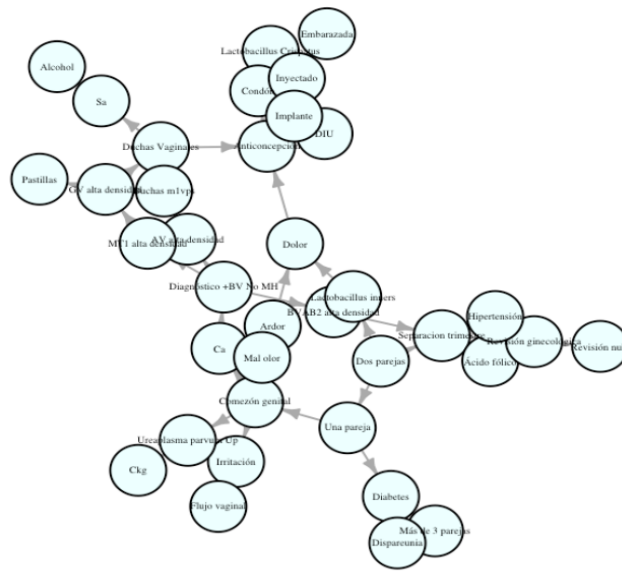


FIGURA 1. Gráfico de Red Bayesiana implementando el algoritmo EM que relaciona todas las variables cualitativas presentadas en el dataset.

En la Figura 2 se visualizan las relaciones de probabilidad creadas por el algoritmo EM en las variables cualitativas del *dataset*, representadas en cuatro columnas. La columna *from* es la primera variable relacionada, la columna *to* muestra con qué variable se está haciendo la relación desde *from* y finalmente las columnas *strength* y *direction* muestran los puntajes de relación de acuerdo a la probabilidad condicional y resultados de los cálculos dados por el algoritmo EM. La Figura 2 son los resultados de probabilidad que tienen las variables expresadas en el gráfico de la Figura 1.

En la Figura 2 se encuentra el porcentaje de relación entre las variables de la columna *from* y *to*. El porcentaje viene representado en las columnas *strength* y *direction*; para la variable AV alta densidad se puede apreciar que tiene 0.985 en los cálculos de probabilidad con respecto a la variable de Diagnóstico +BV, esto indica que se tiene una relación de 98.5 % con respecto a DxBV+.

##	from	to	strength	direction
## 40	Diagnóstico +BV No MH	AV alta densidad	0.985	0.6802030
## 42	Diagnóstico +BV No MH	MT1 alta densidad	0.960	0.7838542
## 43	Diagnóstico +BV No MH	BVAB2 alta densidad	0.775	0.6709677
## 46	Separacion trimestre	Hipertensión	0.615	0.7601626
## 49	Separacion trimestre	Revisión ginecológica	0.545	0.7064220
## 55	Separacion trimestre	Ácido fólico	0.570	0.8201754
## 96	Diabetes	Más de 3 parejas	0.650	0.7769231
## 116	Diabetes	Dispareunia	0.695	0.5647482
## 265	Revisión ginecológica	Revisión nula	1.000	0.8800000
## 353	Una pareja	Dos parejas	1.000	0.8525000
## 426	Dos parejas	Lactobacillus inners	0.520	0.9807692
## 661	Anticoncepción	Condón	1.000	0.8425000
## 663	Anticoncepción	Inyectado	0.995	0.7939698
## 812	Inyectado	Lactobacillus Crispatus	0.595	0.7394958
## 969	Duchas Vaginales	Duchas mlvps	1.000	0.9000000
## 1057	Comezón genital	Ardor	0.715	0.7307692
## 1058	Comezón genital	Irritación	0.650	0.7807692
## 1065	Comezón genital	Ureaplasma parvum Up	0.680	0.7720588
## 1066	Comezón genital	Ca	0.540	0.7129630
## 1102	Ardor	Dolor	0.960	0.8645833
## 1103	Ardor	Mal olor	0.680	0.7867647
## 1147	Irritación	Flujo vaginal	0.750	0.7700000
## 1454	Ureaplasma parvum Up	Ckg	0.845	0.7514793
## 1606	Sa	Alcohol	0.605	0.7107438
## 1744	AV alta densidad	Duchas mlvps	0.750	0.9833333

FIGURA 2. Resultados de la probabilidad condicional determinada como fuerza y dirección de los arcos en las variables cualitativas del dataset.

RESULTADOS

En los resultados finales se realizó la interpretación de los datos. Los resultados fueron dados con base en las experimentaciones de la sección experimentación de los datos anteriormente descrita.

Con respecto a los resultados mostrados en la sección de experimentación, se pudo percatar que todas las variables cualitativas se encuentran relacionadas sin tener una variable objetivo. Después de obtener las tablas de probabilidad de la BN con el algoritmo EM y mostrada en la Figura 2, se elaboraron las relaciones de probabilidad con base en la variable objetivo.

Como resultados finales se buscó encontrar las relaciones que tienen todas las variables cualitativas únicamente con la variable DxBV+. La relación de probabilidad resultante que contiene la relación de las variables cualitativas con la variable DxBV+ y construida con el algoritmo EM se muestra en la Figura 3.

En la Figura 3 se pueden ver la relación que tienen cada una de las variables cualitativas directamente con la variable objetivo. La variable objetivo se encuentra ahora en la columna *from*, las variables cualitativas están enlistadas en la columna *to* y en la columna *strength* y *direction* se cuenta con la probabilidad condicional representada entre estas dos variables.

Con las relaciones mostradas en la Figura 3 se interpreta que la variable AV alta densidad tiene una

##	from	to	strength	direction
## 1	Diagnóstico +BV No MH	Separacion trimestre	0.005	0.0000000
## 2	Diagnóstico +BV No MH	Diabetes	0.000	0.0000000
## 3	Diagnóstico +BV No MH	Hipertensión	0.005	0.5000000
## 4	Diagnóstico +BV No MH	Revisión ginecológica	0.005	0.0000000
## 5	Diagnóstico +BV No MH	Revisión nula	0.000	0.0000000
## 6	Diagnóstico +BV No MH	Una pareja	0.000	0.0000000
## 7	Diagnóstico +BV No MH	Dos parejas	0.000	0.0000000
## 8	Diagnóstico +BV No MH	Más de 3 parejas	0.000	0.0000000
## 9	Diagnóstico +BV No MH	Ácido fólico	0.000	0.0000000
## 10	Diagnóstico +BV No MH	Alcohol	0.000	0.0000000
## 11	Diagnóstico +BV No MH	Anticoncepción	0.080	0.5000000
## 12	Diagnóstico +BV No MH	Condón	0.000	0.0000000
## 13	Diagnóstico +BV No MH	Inyectado	0.015	0.1666667
## 14	Diagnóstico +BV No MH	Duchas Vaginales	0.025	0.0000000
## 15	Diagnóstico +BV No MH	Duchas mivps	0.010	0.0000000
## 16	Diagnóstico +BV No MH	Comezón genital	0.005	0.5000000
## 17	Diagnóstico +BV No MH	Ardor	0.130	0.4615385
## 18	Diagnóstico +BV No MH	Irritación	0.270	0.4722222
## 19	Diagnóstico +BV No MH	Dolor	0.020	0.1250000
## 20	Diagnóstico +BV No MH	Mal olor	0.005	0.0000000
## 21	Diagnóstico +BV No MH	Flujo vaginal	0.005	0.0000000
## 22	Diagnóstico +BV No MH	Ureaplasma parvum Up	0.010	0.5000000
## 23	Diagnóstico +BV No MH	Ca	0.475	0.3842105
## 24	Diagnóstico +BV No MH	Ckg	0.005	0.0000000
## 25	Diagnóstico +BV No MH	Sa	0.000	0.0000000
## 26	Diagnóstico +BV No MH	Lactobacillus Crispatus	0.040	0.2500000
## 27	Diagnóstico +BV No MH	Lactobacillus inners	0.005	0.0000000
## 28	Diagnóstico +BV No MH	AV alta densidad	0.970	0.6958763
## 29	Diagnóstico +BV No MH	GV alta densidad	0.010	0.2500000
## 30	Diagnóstico +BV No MH	MT1 alta densidad	0.965	0.7124352
## 31	Diagnóstico +BV No MH	BVAB2 alta densidad	0.745	0.6812081

FIGURA 3. Resultados de la probabilidad condicional determinada como fuerza y dirección de los arcos en la variable objetivo DxBV+ y las variables cualitativas del dataset.

probabilidad condicional del 97 %, MT1 alta densidad una probabilidad condicional de 96.5 %, BVAB2 alta densidad una probabilidad condicional de 74.5 % pero la bacteria GV alta densidad solo un porcentaje del 0.10 % aunque una relación del 25 % con respecto a las demás variables.

Para la creación del MGC se optó por tomar el algoritmo *Fast Causal Inference* FCI, ya que este algoritmo trabaja con variables discretas o categóricas lo que permitirá en trabajos futuros comparar con otros algoritmos que trabajen con datos cuantitativos. Se siguieron tomando las variables cualitativas.

Después de aplicar el algoritmo EM con los parámetros definidos por defecto, se fijaron los parámetros de búsqueda en un rango de probabilidad condicional mayor al 50 % con respecto a la variable objetivo DxBV+.

En la Figura 4, se usaron las 32 variables mostradas en los datos de la Figura 3, primero se realizó la estructura de la BN con el algoritmo EM y con el resultado obtenido se aplicó el algoritmo FCI.

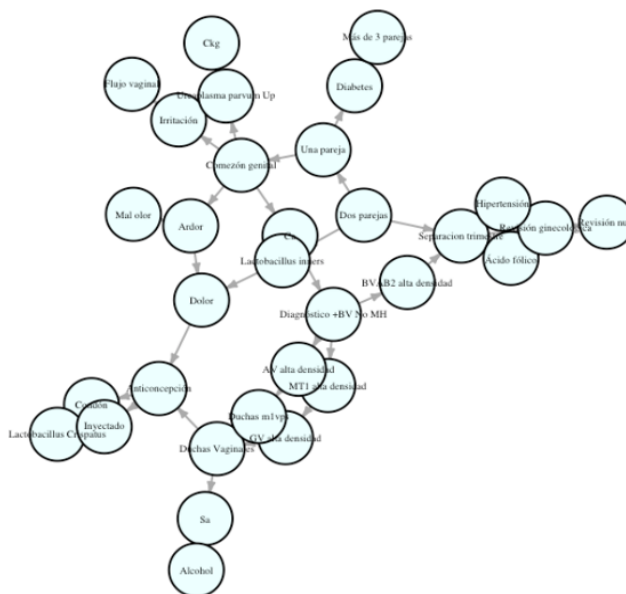


FIGURA 4. Gráfico de BN con algoritmo EM y algoritmo FCI con nodos y arcos dirigidos entre las variables cualitativas y la variable objetivo $DxBV+$.

En la Figura 4 se pueden apreciar nuevamente las relaciones dadas entre las variables cualitativas mediante nodos y arcos. Cuenta con nodos (aristas) y flechas (arcos dirigidos). En esta ocasión se toma como el nodo (A) a la variable Comezón genital y a la variable Ardor como el nodo (B), se interpreta que el nodo A tiene un efecto causal o una influencia directa sobre el nodo B.

Hasta la Figura 4 se tienen relacionadas 32 variables cualitativas que tienen relación directa con el DxBV+, en la siguiente delimitación se reducen las variables que tienen un mayor porcentaje de relación.

El porcentaje de relación para el ejercicio es mayor al 50 % ya que el objetivo es encontrar las variables que tengan una mayor relación con el DxBV+.

De 32 variables iniciales, la experimentación concluye con 22 variables cualitativas incluyendo la variable objetivo. Las 22 variables se encuentran representadas en la BN mostrada en la Figura 5.

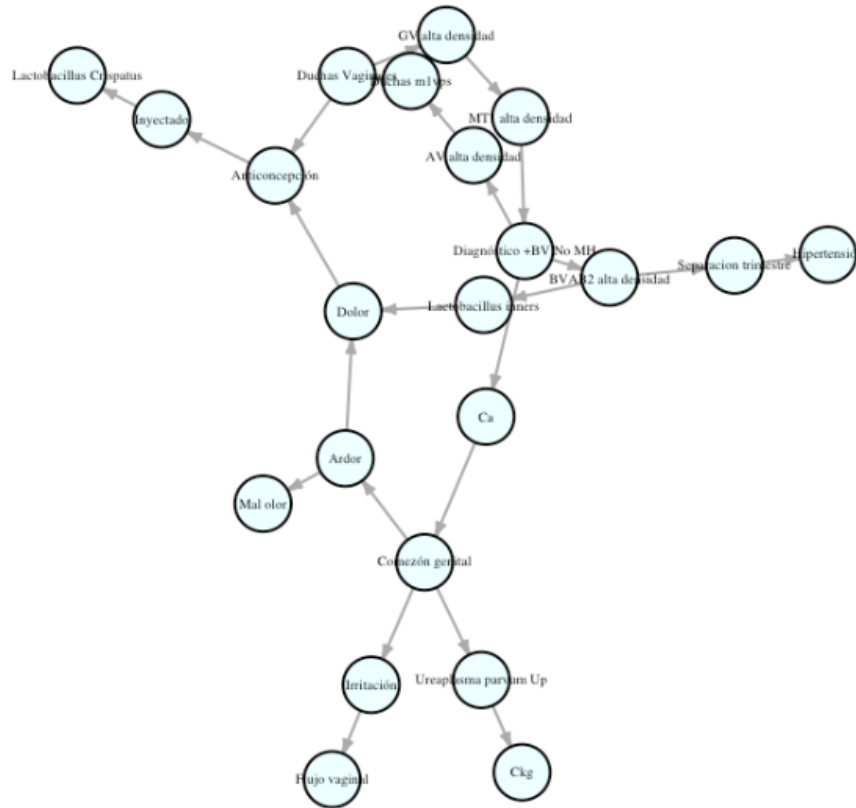


FIGURA 5. Red Bayesiana Causal CBN creada con las 22 variables que tienen relación con el Diagnóstico BV+.

Interpretación y discusión de los resultados

La Figura 5 presenta la CBN en la se tiene la presencia de 22 variables cualitativas que tienen una relación directa con el diagnóstico positivo de la Vaginosis Bacteriana. En la CBN de la Figura 5 se puede apreciar la distribución de los nodos y arcos.

En esta ocasión a manera de ejemplo, se toma como el nodo (A) a la variable MT1 alta densidad y a la variable Diagnóstico +BV para el nodo (B), se interpreta que el nodo A tiene un efecto causal o una influencia directa sobre el nodo B. Otro ejemplo, se puede apreciar en la variable BVAB2 alta densidad nodo (A) con respecto a la variable Lactobacillus inners nodo (B) y también se puede interpretar que la variable BVAB2 alta densidad tiene una influencia directa o un efecto causal sobre la variable

Lactobacillus inners.

La CBN creada se puede seguir explicando nodo por nodo, sin embargo, también es importante mencionar que esta CBN tiene como resultados relaciones de probabilidad condicional. Las relaciones de probabilidad condicional de la CBN muestran el porcentaje de relaciones existentes entre las variables cualitativas que tienen un porcentaje mayor al 50 % con respecto a la variable objetivo Diagnóstico BV+.

##		from	to	strength	direction
## 22	Separacion trimestre	Diagnóstico +BV No MH	0.000	0.0000000	
## 43	Hipertensión	Diagnóstico +BV No MH	0.005	0.5000000	
## 64	Anticoncepción	Diagnóstico +BV No MH	0.120	0.6666667	
## 85	Inyectado	Diagnóstico +BV No MH	0.025	0.7000000	
## 106	Duchas Vaginales	Diagnóstico +BV No MH	0.025	0.7000000	
## 127	Duchas mlvps	Diagnóstico +BV No MH	0.000	0.0000000	
## 148	Comezón genital	Diagnóstico +BV No MH	0.005	0.0000000	
## 169	Ardor	Diagnóstico +BV No MH	0.175	0.5428571	
## 190	Irritación	Diagnóstico +BV No MH	0.310	0.5322581	
## 211	Dolor	Diagnóstico +BV No MH	0.015	0.8333333	
## 232	Mal olor	Diagnóstico +BV No MH	0.000	0.0000000	
## 253	Flujo vaginal	Diagnóstico +BV No MH	0.005	0.5000000	
## 274	Ureaplasma parvum Up	Diagnóstico +BV No MH	0.020	0.5000000	
## 295	Ca	Diagnóstico +BV No MH	0.535	0.5560748	
## 316	Ckg	Diagnóstico +BV No MH	0.000	0.0000000	
## 337	Lactobacillus Crispatus	Diagnóstico +BV No MH	0.040	0.8750000	
## 358	Lactobacillus inners	Diagnóstico +BV No MH	0.000	0.0000000	
## 379	AV alta densidad	Diagnóstico +BV No MH	0.980	0.3520408	
## 400	GV alta densidad	Diagnóstico +BV No MH	0.035	0.7857143	
## 421	MT1 alta densidad	Diagnóstico +BV No MH	0.935	0.3449198	
## 442	BVAB2 alta densidad	Diagnóstico +BV No MH	0.680	0.3161765	

FIGURA 6. Resultados de la probabilidad condicional determinada como fuerza y dirección de los arcos en la variable objetivo DxBV+ y las variables cualitativas finales del dataset.

Finalmente, en la Figura 6 se presentan las relaciones de probabilidad pero de manera inversa a como se presentaron en la Figura 2 y Figura 3. En la columna descrita como *to*, ahora se encuentra la variable objetivo DxBV+, en *from* la variable cualitativa relacionada y finalmente *strength* y *direction* detallan el porcentaje de probabilidad condicional que tienen en relación estas dos variables. Las relaciones de probabilidad se pueden apreciar que las variables mayores a 60 % son AV alta densidad, MT1 alta densidad y BVAB2 alta densidad que son bacterias que de acuerdo con la literatura están relacionadas con el Diagnóstico BV+ [14].

CONCLUSIONES

De acuerdo con la implementación del algoritmo de Red Bayesiana Causal en el *dataset* de Vaginosis Bacteriana en mujeres embarazadas, las variables que se obtuvieron y tienen una relación mayor al 50 % con respecto a la variable del Diagnóstico BV+ fueron:

Anticoncepción, Inyectado, Duchas vaginales, Ardor, Irritación, Dolor, Ca, Lactobacillus crispatus, GV alta densidad.

Por otro lado, las variables que tienen un porcentaje mayor al 60 % en fuerza de correspondencia con respecto a la variable Diagnóstico BV+ fueron:

AV alta densidad, MT1 alta densidad y BVAB2 alta densidad.

Las variables con porcentaje mayor al 60 % coinciden con las que hay en la literatura. El presente estudio muestra que hay bacterias que coinciden con el DxBV+. Algunos estudios señalan que las variables como las duchas vaginales, inyecciones anticonceptivas y calcio también pueden influir en un mínimo porcentaje en el desarrollo de la enfermedad conocida como Vaginosis Bacteriana [15] .

TRABAJOS FUTUROS

Como trabajo futuro se tiene el objetivo de realizar experimentos con otros algoritmos como *tabu search*, *hill climbing* para evaluar el desempeño del algoritmo CBN.

Así mismo se pretende realizar un estudio con variables cuantitativas donde se evalúe el desempeño de Redes Bayesianas Causales desde la perspectiva de los datos cuantitativos del *dataset*.

REFERENCIAS

- [1] J. B. Mina-Ortiz, M. O. Franco-Macias, L. A. Santana, and M. G. Garcia-Ortega. Impacto en la salud materno fetal de embarazadas adolescentes con vaginosis bacteriana. *MQRInvestigar*, 8(1), pp. 5041–5264, 2024.
- [2] Z. Rivera, N. Bravo, R. Borges, and M. Ferreira. Infecciones de transmisión sexual y embarazo. 2010.
- [3] A. A. Lachira Vasquez. Factores asociados a vaginosis bacteriana en mujeres en edad reproductiva en una localidad rural y otra urbano-marginal de Piura–2021. 2023.
- [4] G. Perez Coronel. Conocimiento y actitudes sobre prevención de infecciones de transmisión sexual en estudiantes del Colegio Sagrado Corazón de Jesús, Chota 2023. 2024.
- [5] S. Swidsinski, W. M. Moll, and A. Swidsinski. Bacterial Vaginosis-Vaginal Polymicrobial Biofilms and Dysbiosis. *Deutsches Ärzteblatt International*, 120(20), p. 347, 2023.
- [6] M. Á. Vega, L. M. Q. Mora, and M. V. C. Badilla. Inteligencia artificial y aprendizaje automático en medicina. *Revista médica sinergia*, 5(8), pp. e557–e557, 2020.
- [7] F. Cantú and R. Aldeco. Conocimiento y Razonamiento Computacional. 2019.

- [8] E. M. Rojas. Machine Learning: análisis de lenguajes de programación y herramientas para desarrollo. *Revista Ibérica de Sistemas e Tecnologias de Informação*, (E28), pp. 586–599, 2020.
- [9] G. Soto. El teorema de Bayes. *Revista de Educación Matemática*, 26(3), 2011.
- [10] J. Pearl and D. Mackenzie. El libro del porqué. *La nueva ciencia de la causa y el efecto*. Barcelona, España: Pasado y Presente, 2020.
- [11] L. E. Sucar. Probabilistic graphical models. *Advances in Computer Vision and Pattern Recognition*. London: Springer London, 2015. DOI: 10.1007/978-1-4471-6692-5
- [12] M. Lajous. Inferencia causal en análisis basados en datos de vigilancia epidemiológica para Covid-19. *Salud pública de México*, 63(4), pp. 459–460, 2021.
- [13] K. Train. Métodos de elección discreta con simulación. 2014.
- [14] C. E. Yucra López. Biblioteca para la comparación estadística de algoritmos evolutivos. 2021.
- [15] K. A. M. Morón and C. A. M. Romero. Prevalencia y Factores Asociados con Vaginosis Bacteriana en Universitarias de la Medicina. *Medicina*, vol. 15, p. 15.