

Hacia un Modelo Gráfico Causal para el Análisis de la Vaginosis Bacteriana

Towards a Causal Graphical Model for the Analysis of Bacterial Vaginosis

García-Avalos, M.¹ , Canul-Reich, J.^{1*} , Rodríguez-Henríquez, L.M.² , De la Cruz-Hernández, E.¹ 
(*juana.canul@ujat.mx)

Recibido: 2024-10-30 Aceptado: 2024-11-22

RESUMEN

El Análisis de Trayectorias (*AT*) es una técnica de Estadística Multivariante que estudia los efectos directos (causales) e indirectos de un conjunto de variables observables a través del análisis de correlaciones. Este estudio aplica *AT* a un conjunto de datos sobre la Vaginosis Bacteriana (*VB*) en mujeres embarazadas, el cual tiene tres tipos de diagnósticos. Aunque se han realizado investigaciones sobre las bacterias asociadas a la *VB* utilizando métodos como selección de variables, agrupaciones y reglas de asociación, el *AT* no ha sido utilizado como herramienta visual para detectar las variables coexistentes que varían de una mujer a otra. Se presenta un Modelo Gráfico Causal (*MGC*) que se basa en

un modelo teórico derivado de la matriz de correlación de Pearson y métricas estadísticas. Este modelo muestra las variables que influyen en el diagnóstico de *VB*, identificando las bacterias relevantes: *Mycoplasma Hominis* (*Mh*), *Atopobium Vaginae* (*Av*), *Gardnerella Vaginalis* (*Gv*), *Megasphaera* Tipo 1 (*MT1*) y Bacteria Asociada a Vaginosis Bacteriana Tipo 2 (*BVBA2*). La validez biológica del modelo fue evaluada por un experto en biología. El análisis se llevó a cabo utilizando el software R.

Palabras clave: Vaginosis Bacteriana, Mujeres Embarazadas, Métricas Estadísticas, Correlación, Análisis de Trayectoria.

ABSTRACT

The Trajectory Analysis (*AT*) is a Multivariate Statistical technique that studies the direct (causal) and indirect effects of a set of observable variables through correlation analysis. This study applies *AT* to a dataset on Bacterial Vaginosis (*VB*) in pregnant women, which includes three types of diagnoses. Although research has been conducted on the bacteria associated with *VB* using methods such as variable selection, clustering, and association rules, *AT* has not been utilized as a visual tool to detect

nant women, which includes three types of diagnoses. Although research has been conducted on the bacteria associated with *VB* using methods such as variable selection, clustering, and association rules, *AT* has not been utilized as a visual tool to detect

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

²Instituto Nacional Astrofísica Óptica y Electrónica, Santa María Tonantzintla, Puebla, México, C.P. 72840.

³Universidad Juárez Autónoma de Tabasco, Ranchería sur 4ta secc., Comalcalco Tabasco, C.P. 86650.



coexisting variables that vary from one woman to another. A Causal Graphical Model (*MGC*) is presented, based on a theoretical model derived from the Pearson correlation matrix and statistical metrics. This model illustrates the variables influencing the *VB* diagnosis, identifying the relevant bacteria: *Mycoplasma Hominis* (*Mh*), *Atopobium Vaginae* (*Av*), *Gardnerella Vaginalis* (*Gv*), *Megasphae-*

ra Type 1 (*MT1*), and Bacterial Vaginosis Associated Bacteria Type 2 (*BVBA2*). The biological validity of the model was assessed by a biology expert. The analysis was conducted using R software.

Keywords: Bacterial Vaginosis, Pregnant Women, Statistical Metrics, Correlation, Path Analysis.

INTRODUCCIÓN

La investigación científica en medicina es importante para mejorar el diagnóstico y tratamiento de enfermedades. Se divide en tres tipos: investigación biomédica, clínica y salud pública [1]. La *VB* es común en mujeres en edad fértil, especialmente en las sexualmente activas, y puede ser asintomática o sintomática [2]. El síntoma más destacado en las mujeres es el flujo vaginal anormal, que puede variar en color (gris o verde) y presenta un olor característico a pescado [3]. En biología, se ha observado que múltiples bacterias son responsables de la aparición de la *VB*, y que la composición de estas bacterias puede variar entre mujeres. Este estudio tiene como objetivo identificar las variables coexistentes que influyen en el diagnóstico de la *VB*. Esta enfermedad en el embarazo, influenciada por cambios hormonales, aumenta el riesgo de complicaciones como parto prematuro y endometritis puerperal, por lo que requiere monitoreo y tratamiento [4]. La *VB* puede aumentar la susceptibilidad a infecciones de transmisión sexual, como el Virus del Papiloma Humano (*VPH*). Diversos estudios, tanto transversales como longitudinales, han investigado la relación; sin embargo, no han encontrado asociaciones significativas en estudios transversales entre la *VB* y *VPH* [3].

En el estudio de la *VB* en mujeres sexualmente activas se han utilizado diversos enfoques metodológicos. Uno de ellos es la selección de atributos [5], que identifica las variables más relevantes para predecir un diagnóstico positivo. Otro enfoque es el análisis de agrupaciones [6], que determina grupos de bacterias asociadas con el diagnóstico, observando variaciones entre estos grupos. Además, se han aplicado reglas de asociación para identificar factores coexistentes que contribuyen al estado positivo de la *VB* [7]. Estos métodos buscan aclarar las variables bacterianas que influyen en el diagnóstico de esta condición. Los autores [8] consideran que las tecnologías pueden mejorar los sistemas de salud al analizar grandes datos clínicos para crear modelos predictivos, de tamizaje y diagnóstico. Los gráficos estadísticos [9] son esenciales en el análisis de datos biomédicos, ya que muestran relaciones entre variables. Por otro lado, los Modelos de Ecuaciones Estructurales (*MEE*) permiten comprender las relaciones entre variables observables y no observables. El *AT*, derivado del *MEE*, examina los efectos directos (causales) e indirectos de las variables observables y, mediante un *MGC*, analiza las interacciones entre múltiples variables del sistema [10]. En este trabajo se utilizaron datos de mujeres embarazadas de zonas rurales y urbanas que acudieron a una revisión ginecológica y prueba de diagnóstico de *VB*. Se aplicó el *AT* a estos datos para obtener un *MGC* que ayude a visualizar las variables coexistentes que influyen en el diagnóstico de *VB*.

MATERIALES Y MÉTODOS

Matriz de correlación

Una matriz de correlación es una representación tabular que contiene los coeficientes de correlación entre distintas variables que indica la fuerza y dirección de la relación entre pares de variables. La correlación oscila entre -1 y 1, [11] donde:

- Un valor de 1 indica una correlación positiva perfecta, lo que significa que las variables tienden a aumentar juntas.
- Un valor de -1 indica una correlación negativa perfecta lo que significa que una variable tiende a aumentar, mientras que la otra tiende a disminuir.
- Un valor de 0 indica que no existe correlación entre las variables.

Modelo de Ecuaciones Estructurales

Los MEE son técnicas estadísticas multivariantes para analizar la relación de dependencia o causalidad entre variables observables y no observables. La Figura 1, presenta la nomenclatura utilizada para representar gráficamente un MEE [12].

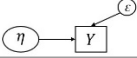
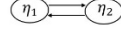
	Rectángulo o Cuadrado: significa que la variable X es observable
	Círculo o Elipse: significa que la variable η es latente
	La variable Y es un indicador de la variable η que tiene un término de error o perturbación ε
	La curva con doble flecha significa asociación entre las dos variables (covariación) ξ_1 y ξ_2
	La presencia de dos flechas simples significa que las variables tienen una causación recíproca

Figura 1: Nomenclatura de Modelo de Ecuaciones Estructurales.

Existen tres categorías de MEE: Análisis de trayectoria, análisis factorial y modelo general de ecuaciones estructurales [12, 13]. El AT examina los efectos causales e indirectos de un grupo de variables observables sobre otras, analizando sus correlaciones. Este análisis se representa gráficamente mediante el MGC, utilizando símbolos y convenciones matemáticas.

Métricas estadísticas de MEE

En estadística, los índices de ajuste (métricas) se utilizan para evaluar el rendimiento de los modelos con datos observados. Los índices miden la calidad del modelo y ayudan a determinar si el modelo describe o predice adecuadamente los datos. La Tabla 1 muestra algunas de las métricas comunes utilizadas en MEE, en particular para AT [10, 14], lo que permite verificar si el modelo teórico propuesto representa adecuadamente la estructura subyacente en los datos.

Los índices *RMSEA* y *SRMR* se conocen como índices absolutos de ajuste y revelan hasta qué punto el modelo representado en la matriz de covarianzas coincide con el modelo implícito. Cuanto menor sea el resultado, mejor será la calidad del ajuste del modelo. Por otro lado, *CFI*, *NFI* y *GFI*, se denominan índices de ajuste incrementales y evalúan cómo mejora un modelo propuesto en comparación con un modelo base [14].

Tabla 1: Métricas Estadísticas

Abreviaturas	Significado	Criterio
RMSEA	Error cuadrático medio de aproximación	< 0.05
SRMR	Raíz cuadrática media normalizada residual	< 0.05
GFI	Índice de bondad de ajuste	≥ 0.95
NFI	Índice de ajuste normalizado	≥ 0.95
CFI	Índice de ajuste comparativo	≥ 0.95

ESTUDIO EXPERIMENTAL

En este estudio se utilizó el lenguaje de programación *R*, que es un software gratuito de análisis estadístico [15], que se ha consolidado como la plataforma estadística utilizada para el análisis de datos. Para el desarrollo del estudio, se inició con la exploración de la estructura de los datos, lo que implicó la eliminación de variables no relacionadas con la VB. Se abordaron también los valores atípicos, que se encontraban fuera del rango normal de cada variable. A continuación, se realizó la imputación de valores, reemplazando los datos faltantes mediante métodos estadísticos. Posteriormente, se calculó la matriz de correlación, obteniendo los coeficientes de correlación entre las variables. Se diseñó un *MGC* que ilustra las relaciones entre las variables, basado en la matriz de correlación, el modelo teórico y las métricas estadísticas. Finalmente, el *MGC* fue validado por personal del laboratorio con experiencia clínica en el área, asegurando su precisión y relevancia.

Conjunto de datos VB

El conjunto de datos se construyó con información de mujeres embarazadas de comunidades rurales y urbanas del estado de Tabasco, sureste de México. Las pacientes se sometieron a un estudio de diagnóstico de VB. Este conjunto de datos incluye información sociodemográfica, bacterias, el diagnóstico de VB, entre otros. La recopilación de los datos se realizó durante las campañas de embarazo saludable (agosto 2018-enero 2020), llevadas a cabo por un investigador biólogo de la División Académica Multidisciplinaria de Comalcalco (*DAMC*) de la Universidad Juárez Autónoma de Tabasco (*UJAT*).

Estructura del conjunto de datos

El análisis exploratorio de datos es un enfoque que utiliza estadísticas descriptivas y herramientas gráficas para obtener una comprensión profunda de un conjunto de datos. Su objetivo es identificar las características del conjunto, como la presencia de valores perdidos y atípicos, los cuales deben ser abordados para mejorar la calidad y la validez de los resultados obtenidos en el análisis posterior [16]. En el análisis exploratorio de datos se identificaron 87 atributos y 132 instancias, de los cuales 41 eran atributos categóricos y 46 atributos numéricos. El atributo clase (Diagnóstico de Vaginosis Bacteriana sin *Mycoplasma Hominis*, *DxVBNoMh*) se dividió en tres categorías: Vaginosis Bacteriana Positivo (*VB+*) con 32 instancias, que indica la presencia de la condición; Indeterminado con 5 instancias, para casos sin un diagnóstico claro; y Microbiota Normal (*VB-*) con 95 instancias, que señala la ausencia de la condición. Como se observa, el conjunto de datos está desbalanceado con *VB-* como clase mayoritaria. El impacto del desbalanceo de clases no forma parte de este estudio, sin embargo, la investigación está en curso y ello corresponde a una etapa posterior. Se registraron 609 valores faltantes y 143 valores atípicos en el conjunto de datos.

Eliminación de atributos

Durante el análisis se identificaron variables que no están relacionadas con la VB, esto se verificó con el área clínico-biológica quien fue creador de los datos. Al tener este conocimiento se decidió eliminar tales variables, en particular se descartó todas las variables relacionadas con el VPH. Así, de los 87 atributos originales, se obtuvo 72 variables con 132 observaciones.

Valores atípicos

El diagrama de caja fue desarrollado por Tukey en 1977 como una herramienta para el análisis exploratorio de datos. Se utiliza para resumir y comparar distribuciones basándose en sus valores extremos (máximo y mínimo), así como en la mediana y los cuartiles. Este gráfico también se conoce como diagrama de caja y bigotes [17]. Se implementó la técnica estadística en el conjunto de datos para identificar los valores atípicos, resultando en la detección de 143 valores atípicos. A pesar de que estos datos se encuentran fuera del rango normal en las variables, corresponden a pacientes que presentan otras enfermedades. Tras consulta con un biólogo experto en el área, se determinó que era pertinente mantener estos valores originales en el conjunto de datos. En la Figura 2, se muestran diagramas de cajas de algunos atributos del conjunto de datos. En la gráfica, se observa la mediana de estas variables y se resalta que esta técnica permite identificar los valores atípicos, los cuales se extienden más allá de la longitud del bigote, como en los casos de los atributos: InicioSexual, N.EMBARAZO y N.DEHIJOS.

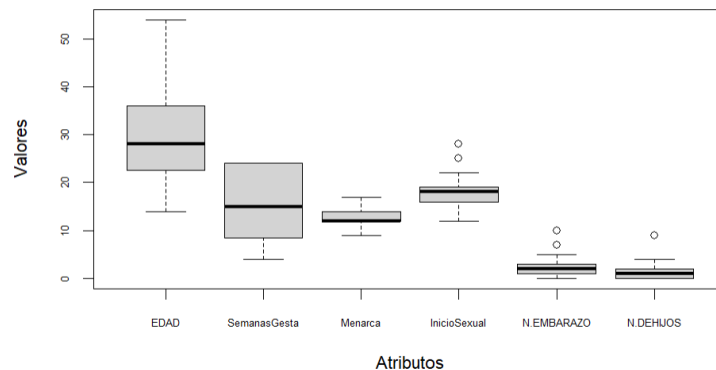


Figura 2: Diagrama de caja de los primeros atributos del conjunto de datos.

Valores faltantes

La imputación de datos faltantes es el proceso que utiliza la información disponible en la muestra para asignar un valor a las variables con datos ausentes. Su objetivo es crear un conjunto de datos completo para el análisis que se desea realizar. La imputación por media, que consiste en utilizar el promedio de una variable, es una técnica estadística aplicada para manejar valores faltantes en variables numéricas [18]. Por otro lado, la imputación por moda, que se refiere a la frecuencia más común de una variable, es una técnica estadística empleada para abordar valores ausentes en variables categóricas [19]. En el análisis se detectaron 609 datos perdidos distribuidos en 47 variables, que correspondían a atributos tanto categóricos como numéricos. Para imputar los datos faltantes en las variables categóricas utilizando la moda, primero se identificaron los valores faltantes de cada atributo y se determinó el diagnóstico (VB+, Indeterminado o VB-) correspondiente. Luego, se agruparon las instancias completas de cada atributo según el diagnóstico, se calculó la moda de cada grupo y, finalmente, se reemplazaron los datos ausentes

por las modas asociadas a su diagnóstico. Por ejemplo, en la Tabla 2 se muestran los datos imputados de las primeras 10 instancias de la variable uso de anticonceptivo (Anticoncepción). La columna *DxVBNoMh* indica el diagnóstico de VB, donde 1 representa VB+, 2 indica indeterminado y 3 corresponde a VB-.

Tabla 2: Datos imputados con la moda del atributo categórico de Anticoncepción y diagnóstico de VB

Instancias	Datos originales	Datos imputados	DxVBNoMh
1	NA	2	3
2	NA	2	1
3	NA	2	3
4	NA	2	3
5	2	2	3
6	NA	2	3
7	2	2	3
8	2	2	3
9	2	2	3
10	1	1	3

Para imputar los datos perdidos en las variables numéricas con la media, se siguió un proceso similar a los atributos categóricos, descrito en el párrafo anterior. La Tabla 3 presenta los datos imputados de las primeras 10 instancias del atributo inicio sexual (InicioSexual).

Tabla 3: Datos imputados con media del atributo numérico InicioSexual y diagnóstico de VB

Instancias	Datos originales	Datos imputados	DxVBNoMh
1	NA	18.0137	3
2	NA	17.5000	1
3	NA	18.0137	3
4	NA	18.0137	3
5	NA	18.0137	3
6	NA	18.0137	3
7	NA	18.0137	3
8	NA	18.0137	3
9	NA	18.0137	3
10	NA	18.0137	3

Matriz de correlación de Pearson

Para determinar las relaciones entre las variables y el atributo de clase (*DxVBNoMh*), dada la cantidad de atributos en el conjunto de datos y con el fin de visualizar la matriz de correlación con claridad, los datos se agruparon en bloques pequeños atendiendo el orden en que se encuentran en el conjunto de datos de manera secuencial y como criterio único mantener un tamaño equilibrado en cada grupo. Se

calculó la matriz de correlación de Pearson. La Figura 3 presenta las matrices de correlación de dos de los grupos formados. En la matriz de la izquierda se resaltan cuatro variables (Av , Gv , $MT1$ y $BVAB2$) con coeficientes de correlación significativos, mientras que en la matriz de la derecha solo se observa una variable (Mh) relevante.

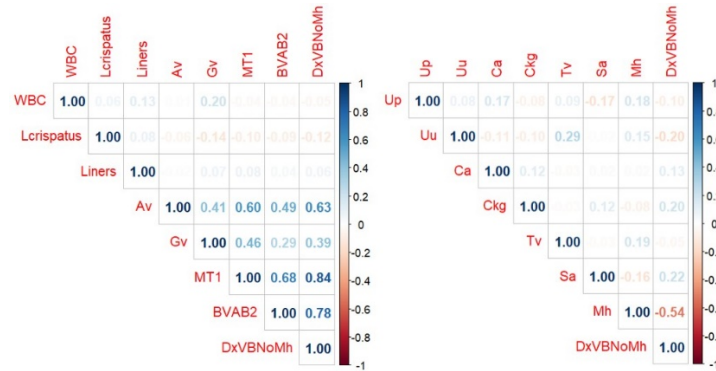


Figura 3: Conjuntos de matrices de correlación.

Después de calcular las matrices de correlación para cada grupo, se seleccionaron las variables con correlaciones negativas ≤ -0.54 y positivas ≥ 0.39 respecto al atributo "de clase". Aunque los coeficientes de correlación son bajos, se consideran relevantes debido al contexto clínico y en nuestro caso con datos reales [20, 21]. Los coeficientes identificados fueron 0.39, 0.63, 0.78, 0.84 y -0.54. Se formó un grupo que incluye todas las variables que cumplen con estos criterios de coeficiente de correlación. La Figura 4 muestra la matriz de correlación triangular superior de las variables seleccionadas (Mh , Av , Gv , $MT1$, $BVAB2$) y el $DxVBNomh$. La diagonal principal de la matriz, que va de la esquina superior izquierda a la esquina inferior derecha, está formada por unos, esto indica una correlación perfecta de una variable consigo misma.

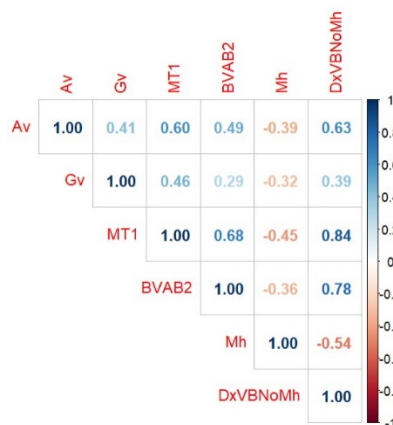


Figura 4: Matriz de correlación de los atributos relevantes.

Modelo Gráfico Causal

El desarrollo de un modelo gráfico causal requiere la determinación del modelo teórico basado en la matriz de correlación. Se inicia presentando modelos teóricos simples, que involucran únicamente una variable y el atributo de clase. Los modelos teóricos simples Ec. (1), Ec. (2), Ec. (3), Ec. (4) y Ec. (5) se representan en los diagramas a), b), c), d) y e), de la Figura 5, respectivamente.

$$Mod1 \leftarrow' DxVBNoMh \sim Av' \quad (1)$$

$$Mod2 \leftarrow' DxVBNoMh \sim Gv' \quad (2)$$

$$Mod3 \leftarrow' DxVBNoMh \sim MT1' \quad (3)$$

$$Mod4 \leftarrow' DxVBNoMh \sim BVAB2' \quad (4)$$

$$Mod5 \leftarrow' DxVBNoMh \sim Mh' \quad (5)$$

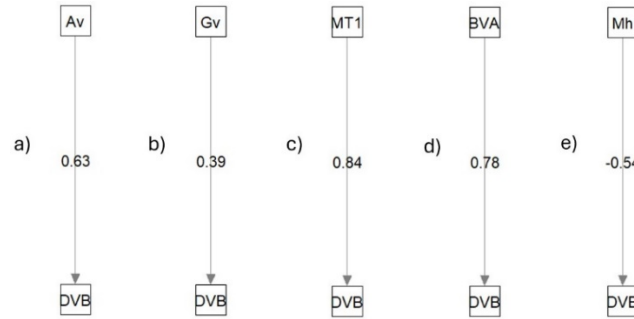


Figura 5: Modelos gráficos causales simples. Abreviaturas: BVA representa BVAB2 y DVB representa DxVBNoMh.

Para evaluar el ajuste de los *MGC* simples Ec. (1), Ec. (2), Ec. (3), Ec. (4) y Ec. (5), se calcularon las métricas estadísticas presentadas en la Tabla 1. Los valores obtenidos de las métricas para los cinco modelos teóricos son: $RMSEA = 0.000$, $SRMR = 0.000$, $GFI = 1.000$, $NFI = 1.000$ y $CFI = 1.000$, todos dentro de los rangos esperados de acuerdo con la Tabla 1. Los coeficientes de trayectoria deben interpretarse de la siguiente manera: los valores cercanos a ± 1 indican una fuerte asociación entre las variables o con el atributo, mientras que los valores cercanos a 0 reflejan una asociación débil. La Figura 5 muestra las relaciones causales entre las variables *Av*, *Gv*, *MT1*, *BVA* y *Mh* respecto al *DVB*, con los coeficientes de las trayectorias (0.63, 0.39, 0.84, 0.78 y -0.54) obtenidos a partir de los *MGC* simples. Estos coeficientes reflejan las relaciones según los modelos teóricos generados, y coinciden con los coeficientes de correlación de Pearson. Los coeficientes de trayectoria positivos indican que, a medida que aumentan las bacterias, también incrementa la probabilidad de obtener el *DVB*. En contraste, el coeficiente de trayectoria negativo indica que la reducción de las bacterias está asociada con un aumento en la probabilidad de tener el *DVB*.

Después de analizar los *MGC* simples, se aplica el mismo *AT* para desarrollar un *MGC* que incluye las variables relevantes junto con el atributo de clase, con el fin de visualizar las relaciones que se determinan al basarnos en la matriz de correlación de la Figura 4. La obtención del modelo teórico Ec. (6), se construyó de la siguiente forma: se comenzó considerando el *DxVBNoMh* en relación con las variables predictoras *Mh*, *MT1*, *BVAB2*, *Av* y *Gv*, es decir la primera relación. Luego, se crearon las demás relaciones basándose en la variable *MT1* (por estar más correlacionada con el diagnóstico de *VB*) y las otras variables restantes, respecto al atributo clase.

$$ModC \leftarrow' DxVBNoMh \sim Mh + Av + Gv + MT1 + BVAB2$$

$$MT1 \sim Mh$$

$$MT1 \sim Av$$

$$MT1 \sim Gv$$

$$MT1 \sim BVAB2' \quad (6)$$

A medida que se añadían relaciones, se calculaban las métricas estadísticas para verificar que se cumplieran las condiciones enumeradas en Tabla 1. Los valores obtenidos de las métricas del modelo Ec. (6), que abarca todos los modelos teóricos, son: $RMSEA = 0.000$, $SRMR = 0.000$, $GFI = 1.000$, $NFI = 1.000$ y $CFI = 1.000$, estos valores indican que el modelo teórico está dentro de los rangos esperados de la Tabla 1. El modelo Ec. (6) permitió desarrollar el *MGC* mostrado en la Figura 6, que ilustra las relaciones causales o directas (líneas continuas) entre todas las variables y el *DVB*, así como las covariaciones (líneas discontinuas). Las relaciones causales desde las variables *Mh*, *Av*, *Gv*, *BVA* y *MT1* hacia *DVB* presentan coeficientes de -0.16, 0.14, -0.04, 0.35 y 0.46, respectivamente. Además, se identifican relaciones indirectas, como la conexión entre *BVA* y *DVB* a través de *MT1*, donde *MT1* actúa como variable intermedia, con coeficientes de 0.46.

Validación biológica del MGC

El *MGC* es una herramienta visual que ilustra las relaciones entre las variables y el *DVB*, facilitando la identificación de las bacterias coexistentes *Mh*, *Av*, *Gv*, *BVA* y *MT1* que influyen en el *DVB*, las cuales son frecuentes según la literatura [22, 23]. El *MGC* indica que la bacteria *Gv* tiene la más baja asociación causal con el *DVB*, con un coeficiente de -0.04. No obstante, *Gv* muestra una mejor relación con otras bacterias relacionadas con el *DVB*. Por otro lado, la bacteria *MT1* presenta una fuerte asociación causal con el *DVB*, con un coeficiente de 0.46.

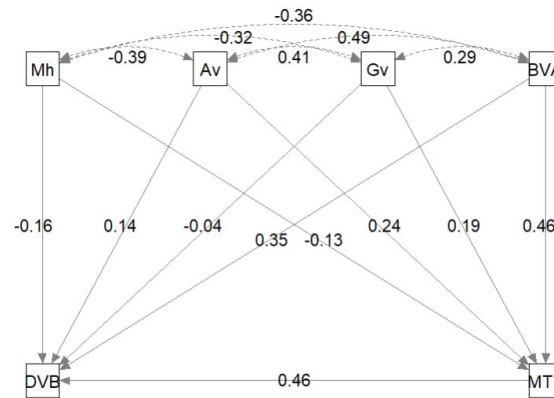


Figura 6: Modelos gráficos causal de las variables relevantes.

RESULTADOS Y DISCUSIÓN

En este estudio se aplicó el método de *AT* debido a la presencia de variables observables en el conjunto de datos de mujeres embarazadas con *DVB*, que incluye 87 variables y 132 observaciones. El atributo de clase abarca tres categorías: Vaginosis Bacteriana con 32 instancias, Indeterminado con 5 instancias y Microbiota Normal con 95 instancias. Se trabajó con 72 variables y 132 instancias (en la sección de valores faltantes se muestran las primeras 10 instancias). La matriz de correlación del conjunto permitió identificar relaciones significativas entre las variables y el atributo de clase, lo que facilitó la determinación de los modelos Ec. (1) a Ec. (6). Estos modelos fueron evaluados mediante métricas estadísticas, contribuyendo a la construcción de modelos gráficos causales ilustrados en las Figuras 5 y 6. Las variables obtenidas en el análisis de correlación son *MT1*, *BVA*, *Av*, *Gv* y *Mh*, las cuales se visualizan en los *MGC* con respecto a la condición. Además, se muestran las relaciones causales e indirectas entre las variables y el diagnóstico, así como los coeficientes de trayectorias. Los resultados presentados fueron validados biológicamente por el experto por lo que la interpretación del modelo desde el campo biológico permite ofrecer un modelo causal confiable construido con datos reales de pacientes. La investigación está en curso y el modelo aún es preliminar, por lo que el diseño experimental de la siguiente fase permitirá ampliar y dar mayor soporte para utilizar el *MGC* como un modelo causal que identifique variables detonantes de la *VB*.

CONCLUSIONES

En este estudio, se utilizó el método de *AT* debido a que el conjunto de datos de mujeres embarazadas con el *DVB* consta de variables observables. Se trabajó con 72 variables y 132 instancias. El *AT* se aplicó a partir del análisis de correlación, lo que permitió identificar relaciones significativas con el atributo de clase y facilitar la formulación de modelos evaluados mediante métricas estadísticas. Las variables obtenidas en el análisis de correlación *MT1*, *BVA*, *Av*, *Gv* y *Mh* se visualizaron en los *MGC* en relación con el *DVB*. Además, se mostraron las relaciones causales e indirectas entre las variables y *DVB*, así como los coeficientes de trayectoria. A diferencia de los enfoques tradicionales, el *MGC* permite detectar tanto asociaciones fuertes como débiles, revelando interacciones que no son evidentes en otros modelos. Un hallazgo importante es que la literatura ha identificado a *MT1* como una bacteria fuertemente asociada al diagnóstico de *DVB*, lo cual es confirmado con el *MGC*. Además de corroborar lo que dice la literatura,

en el *MGC* se puede observar las correlaciones entre las bacterias mismas y no solo de las bacterias con el *DVB*. El *MGC*, como herramienta visual, podría facilitar en el futuro la identificación de las bacterias más influyentes en el desarrollo de la *VB*, lo que permitiría mejorar su diagnóstico y tratamiento. Adicionalmente, con la interpretación presentada del *MGC*, se sientan las bases para aplicarlo en el análisis causal de la *DVB* y para otras enfermedades.

DECLARACIÓN DE CONTRIBUCIÓN DE AUTORES Y COLABORADORES

Maricela García: Ejecución de experimentos y redacción. **Juana Canul:** Diseño experimental y supervisión. **Lil María Rodríguez:** Diseño experimental y revisión. **Erick De la Cruz:** Validación biológica de los hallazgos.

REFERENCIAS

- [1] D. Mendoza, A. Quintana, y A. Díaz. La importancia de la investigación científica en medicina. *Revista Médica Sinergia, Atención Familiar UNAM*, vol. 24, no. 3, pp. 224–227, 2023. DOI: 10.22201/fm.14058871p.2023.3.85789.
- [2] C. Celeste *et al.* Ethnic disparity in diagnosing asymptomatic bacterial vaginosis using machine learning. *npj Digital Medicine*, vol. 6, no. 211, 2023. DOI: 10.1038/s41746-023-00953-1.
- [3] E. Sánchez *et al.* Molecular epidemiology of bacterial vaginosis and its association with genital micro-organisms in asymptomatic women. *Journal of Medical Microbiology*, vol. 68, no. 9, pp. 1373–1382, 2019. DOI: 10.1099/jmm.0.001044.
- [4] I. Morelli y S. Gamboa. Vaginosis bacteriana en el embarazo: últimos avances hasta la fecha. *Revista Médica Sinergia*, vol. 7, no. 7, pp. e838, 2022. DOI: 10.31434/rms.v7i7.838.
- [5] J. Pérez, J. Canul, J. Hernández, y B. Hernández. Selección de predictores para el diagnóstico de la vaginosis bacteriana utilizando árboles de decisión y algoritmos de relief. *Applied Sciences*, vol. 10, no. 9, pp. 3291, 2020. DOI: 10.3390/app10093291.
- [6] H. Hernández, J. Canul, B. Hernández, and E. De la Cruz. An agglomerative hierarchical clustering approach to identify coexisting bacteria in groups of bacterial vaginosis patients. *Intelligent Data Analysis*, vol. 27, no. 3, pp. 583–611, 2023. DOI: 10.3233/IDA-216488.
- [7] F. De la Cruz, J. Canul, R. Rivera, and E. De la Cruz. Impact of data balancing a multiclass dataset before the creation of association rules to study bacterial vaginosis. *Intelligent Medicine, Elsevier*, vol. 4, no. 3, pp. 188–199, 2023. DOI: 10.1016/j.imed.2023.02.001.
- [8] M. Álvarez, L. Quirós, y M. Cortés. Inteligencia artificial y aprendizaje automático en medicina. *Revista Médica Sinergia*, vol. 5, no. 8, pp. 1–12, 2020. DOI: 10.31434/rms.v5i8.557.
- [9] N. Sagaró y L. Zamora. Métodos gráficos en la investigación biomédica de causalidad. *Revista Electrónica Dr. Zoilo E. Marinello Vidaurreta*, vol. 44, no. 4, 2019. Disponible en: <http://revzoilomarinellosldcu/index.php/zmv/article/view/1846>.

-
- [10] A. Lepera. Introducción a los Modelos de Ecuaciones Estructurales y su Implementación en R mediante un ejemplo. *Revista de Investigación en Modelos Matemáticos Aplicados a la Gestión y la Economía*, vol. 1, no. 8, pp. 15–37, 2021. Disponible en: <https://www.economicas.uba.ar/wp-content/uploads/2016/04/Lepera-Andrea-1.pdf>.
- [11] M. Molinas, C. Ochoa, y E. Ortega. Correlación, Modelos de regresión, Evidencia Pediátrica. vol. 17, no. 2, pp. 17–25, 2021. Disponible en: <https://evidenciasenpediatria.es/articulo/7827/correlacion-modelos-de-regresion>.
- [12] J. Iglesias. Modelo de Ecuaciones Estructurales. Universidad de Almería, pp. 1–72, 2021. Disponible en: <https://repositorio.ual.es/bitstream/handle/10835/13177/IGLESIAS%20LABRACA,%20JUAN%20MIGUEL.pdf?sequence=1>.
- [13] K. Barbeau, K. Boileau, F. Sarr, and K. Smith. Path analysis in Mplus: A tutorial using a conceptual model of psychological and behavioral antecedents of bulimic symptoms in young adults. *The Quantitative Methods for Psychology*, vol. 15, no. 1, pp. 38–53, 2019. DOI: 10.20982/tqmp.15.1.p038.
- [14] F. Jordan. Valor de corte de los índices de ajuste en el análisis factorial confirmatorio. *Revista de Investigación en Psicología Social, Redalyc*, vol. 7, no. 1, pp. 66–71, 2021. Disponible en: <https://www.redalyc.org/articulo.oa?id=672371335005>.
- [15] F. Jahuey, J. Herrera, y F. Paredes. El programa R: una estrategia inicial para su entendimiento y aprendizaje. *Revista Digital Universitaria*, vol. 23, no. 4, 2022. DOI: 10.22201/cuaieed.16076079e.2022.23.4.4.
- [16] C. Bouza. Análisis exploratorio de datos univariados para la ciencia de datos. *Red Iberoamericana de Estudios Cuantitativos Aplicados*, 2023. DOI: 10.13140/RG.2.2.26038.06722.
- [17] A. Madrid, S. Valenzuela, C. Batanero, y J. Garzón. Interpretación del diagrama de caja por estudiantes universitarios de ciencias de la actividad física y deporte. *Educación Matemática*, vol. 34, no. 3, 2022. DOI: 10.24844/EM3403.10.
- [18] A. Contreras, N. Luna, P. Macías, y N. González. Imputación de datos faltantes en encuestas en hogares. *Memorias de la Décima Cuarta Conferencia Iberoamericana de Complejidad, Informática y Cibernética*, pp. 310–316, 2024. DOI: 10.54808/CICIC2024.01.310.
- [19] E. Mejía, J. Aguilar, P. Bolaños, y J. López. Métodos de imputación por regresión, imputación por moda, imputación múltiple y árboles de decisión para variables categóricas en perspectiva comparada. *Revista De La Universidad Del Zulia*, vol. 15, no. 43, pp. 541–555, 2024. DOI: 10.46925//rd-luz.43.29.
- [20] L. Poaquiza y G. Gavilanes. El estrés y su relación con la inteligencia emocional en docentes universitarios. *Ciencia Latina Revista Científica Multidisciplinar*, vol. 6, no. 3, pp. 1357–1372, 2022. DOI: 10.37811/cl_rcm.v6i3.2301.

- [21] R. Ponce y M. Cáceres. Dependencia emocional y su relación con el riesgo suicida en adultos jóvenes. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades*, vol. 4, no. 1, pp. 329–339, 2023. DOI: 10.56712/latam.v4i1.247.
- [22] R. Drew *et al.* An interpretation algorithm for molecular diagnosis of bacterial vaginosis in a maternity hospital using machine learning: proof-of-concept study. *Diagnostic Microbiology and Infectious Disease*, vol. 96, no. 2, pp. 114950, 2020. DOI: 10.1016/j.diagmicrobio.2019.114950.
- [23] J. Salas, L. Angulo, y E. Garita. Vaginosis Bacteriana-Actualización y novedad terapéutica. *Revista Ciencia y Salud, UCIMED*, vol. 5, no. 6, pp. 85–93, 2022. DOI: 10.34192/cienciaysalud.v5i6.387.