Universidad Juárez Autónoma de Tabasco

División Académica de Ciencias y Tecnologías de la Información

Tesis doctoral

“Identificación de predictores para el diagnóstico de Vaginosis Bacteriana”

Que presenta

Jesús Francisco Pérez Gómez

Para obtener el grado de

Doctor en Ciencias de la Computación

Director

Dra. Juana Canul Reich

Dr. José Hernández Torruco

Liga de Generación y Aplicación del Conocimiento:
Ciencia de Datos e Inteligencia Artificial

Cunduacán, Tabasco, México Enero 2022

Universidad Juárez Autónoma de Tabasco

División Académica de Ciencias y Tecnologías de la Información

Tesis doctoral

Identificación de predictores para el diagnóstico de Vaginosis Bacteriana

Que presenta

Jesús Francisco Pérez Gómez

Para obtener el grado de

Doctor en Ciencias de la Computación

Comité tutoral

Dra. Juana Canul Reich
Dr. José Hernández Torruco

Jurado

Dra. Betania Hernández Ocaña
Dr. Erick Natividad Hernández de la Cruz
Dr. Rafael Rivera López
Dra. Juana Canul Reich
Dr. José Hernández Torruco

Cunduacán, Tabasco, México Enero 2022

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

F5: Liberación de dirección de tesis

Cunduacán, Tabasco., a 17 de enero de 2022

MTE. Oscar Alberto González González
Director de la División Académica de Ciencias y Tecnologías de la Información
Presente

Por medio de la presente nos permitimos comunicarle que después de haber concluido la dirección de la Tesis: ***"Identificación de predictores para el diagnóstico de Vaginosis Bacteriana"***, elaborada por el ***C. Jesús Francisco Pérez Gómez***, del Doctorado en Ciencias de la Computación, consideramos que puede continuar con los trámites para la obtención del grado.

Sin otro particular, aprovechamos la ocasión para enviarle un cordial saludo.

A t e n t a m e n t e

Dra. Juana Canul Reich

Dr. José Hernández Torruco

c.c.p. Dr. Eddy Arquímedes Ancona Alcocer. Encargado del despacho de la Coordinación de Posgrado.
Directores de Tesis.
Estudiante

F6: Solicitud de Jurado 2022

Cunduacán, Tabasco, a 19 de enero de 2022.

Asunto: Solicitud de Jurado

MTE. Oscar Alberto González González
Director de la DACYTI
Presente

Por este medio me permito informarle que la tesis: "Identificación de predictores para el diagnóstico de vaginosis bacteriana", ha sido liberada por nuestros asesores: Juana Canul Reich y José Hernández Torruco, por lo que en atención a ello me dirijo a usted con la finalidad de solicitarle tenga a bien nombrar al jurado para que evalúe el citado trabajo.

Sin otro particular, aprovecho la ocasión para enviarle un cordial saludo.

Atentamente

Jesus Francisco Pérez Gómez

Matrícula:	181H18002
Domicilio:	Leovigildo Leyva SN, Col. El Carmen
Localidad:	Nacajuca, Tabasco
Teléfono:	9931594577
E-mail:	iscfranciscoperez@gmail.com

c.c.p. Dr. Eddy García Alcocer. - Coordinador de posgrado.
Interesado.

F7: Respuesta de Jurado

2022

Cunduacán, Tabasco., a 22 de enero de 2022.

Asunto: Respuesta de Jurado

MTE. Oscar Alberto González González
Director de la DACYTI
Presente

En atención a los oficios girados por usted, en los que se nos designa como parte del jurado para efectuar la revisión de la tesis titulada ***"Identificación de predictores para el diagnóstico de vaginosis bacteriana"***, realizada por el C. Jesús Francisco Pérez Gómez, estudiante del Doctorado en Ciencias de la Computación, nos permitimos informarle que en virtud de que ha atendido las observaciones realizadas, otorgamos nuestra aprobación para que continúe los trámites correspondientes a la obtención del grado.

Sin otro particular, aprovechamos la ocasión para enviarle un cordial saludo.

Atentamente integrantes del Jurado

Dra. Betania Hernández Ocaña

Dr. Erick Natividad De La Cruz Hernandez

Dr. José Adán Hernández Nolasco

C.c.p. Dr. Eddy García Alcocer. - Coordinador de posgrado.
Integrantes del Jurado
Estudiantes

"2022, Año de Ricardo Flores Magón"

Cunduacán, Tabasco a 19 de enero de 2022
Oficio No. 010/DACYTI/CP/2022

Asunto: Autorización de Modalidad de examen

C. Jesús Francisco Pérez Gómez
Doctorado en Ciencias de la Computación

En atención al oficio de fecha 19 de enero de 2022, en el cual solicita la autorización para titularse bajo la modalidad de Tesis, me permito informarle que se **AUTORIZA** el examen de grado, en virtud de que cumple satisfactoriamente los requisitos establecidos en el Reglamento General de Estudio de Posgrado vigente en la Universidad y los establecidos por el plan de estudios del Doctorado en Ciencias de la Computación

Sin otro particular, aprovecho la ocasión para saludarle cordialmente.

Atentamente

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

M.T.E. Óscar Alberto González González
Director

C.c. p. Dr. Eddy Arquímedes García Alcocer. – Encargado del despacho de la Coordinación de Posgrado.
Archivo.
Consecutivo.

M.T.E. OAGG/EAGA

F8: Cesión de Derechos 2022

Cunduacán, Tabasco., a 22 de enero de 2022.

Asunto: Cesión de Derechos.

A quien corresponda:

Los que suscriben la presente, declaramos que el proyecto de titulación denominado, **"Identificación de predictores para el diagnóstico de vaginosis bacteriana"** es de nuestra autoría intelectual y por lo tanto cedemos todos los **derechos** sobre este proyecto a la Universidad Juárez Autónoma de Tabasco, a la cual relevamos de cualquier sanción y asumimos responder a cualquier reclamo de derechos de autor ante las autoridades competentes.

Atentamente

Autores:

Nombre	Domicilio	Firma autógrafa
MATI. Jesús Francisco Pérez Gómez	Leovigildo Leyva SN, Colonia el Carmen, Nacajuca, Tabasco. CP86220.	
Dra. Juana Canul Reich	Rinconada del Vergel 130, Fraccionamiento Hacienda Esmeralda, Villahermosa, Tabasco. CP86144	
Dr. José Hernández Torruco	Pedro Gutiérrez Cortés 406 Col. Gaviotas Norte, Villahermosa, Centro, Tabasco. CP86090	

c.c.p. Oscar González González. - Director de la DACYTI
Ccp. Interesados

Universidad Juárez Autónoma de Tabasco

Secretaría de Servicios Académicos

Dirección de Servicios Escolares

181H18002

Villahermosa, Tabasco, 21 de enero de 2022

Alumno: **Jesús Francisco Pérez Gómez**

Matrícula: 181H18002 **CURP:** PEGJ861001HTCRMS09

Posgrado: Doctor en Ciencias de la Computación

Antecedentes Académicos

Estudios de Posgrado de Maestro en Administración de Tecnologías de la Información

Entidad: Tabasco

Período: 2013 - 2015 **Cédula:** 9977824

Institución: Universidad Juárez Autónoma de Tabasco

Validó y Cotejó: MDL00527 - 14 de mayo de 2018

Se autoriza examen profesional en la modalidad de tesis

Hago constar que he verificado la información aquí presentada (antecedentes académicos y datos personales anexos), la cual es fidedigna y autorizo a la Universidad para que haga uso de ella para la emisión de título electrónico y la representación impresa del mismo.

Jesús Francisco Pérez Gómez

Nombre y firma del alumno

LA. Marciana León Ficachi
Autoriza

Dedicatoria

Principalmente dedico esta tesis a quien hace posible que hoy en día me encuentre con vida, con salud, y, sobre todo, a quien me otorgó una familia tan maravillosa como la que tengo, a mi padre Dios.

De la mima manera dedico este trabajo a las personas que, aunque no estuvieron desvelándose, escribiendo ni leyendo junto a mí, siempre han estado conmigo en las buenas y en las malas:

A mi Madre, pues desde que era pequeño me ha sabido inculcar a mí y a mi hermana las herramientas claves para la superación: el maravilloso hábito del estudio.

A mi esposa, que sin lugar a dudas es mi apoyo moral, quien ha sabido entender y comprender de primera mano todas aquellas noches, días y horas que tuve que invertir para llegar a alcanzar ésta gran meta de mi vida.

A mi Padre, quien me ha enseñado que no existe mayor satisfacción que el dedicarse a lo que uno le gusta hacer, pues con su ejemplo de dedicación y responsabilidad ha logrado cosechar esos frutos del esfuerzo.

A mis hijos Paquito, Xime y Vale, quienes con su alegría, entusiasmo y creatividad han sabido inyectarme gran júbilo a mi vida, pues me llenan con su bondad.

Agradecimientos

En primer lugar, quiero agradecer a la máxima casa de estudios de los tabasqueños, la UJAT, quien ha sido mi segundo hogar en estos años de esfuerzo y dedicación, y quien me ha permitido aprovechar sus espacios para realizar y fomentar las actividades realizadas en este trabajo final.

Un agradecimiento especial a mis directoras de tesis Dra. Juana Canul y el Dr. José Torruco por compartir este tiempo de aprendizaje y enseñanza, quienes sé que con mucho gusto me llevaron de la mano en el andar académico. Estoy seguro que gracias a ellos, más que una investigación excepcional, este trabajo ha finalizado en una amistad inquebrantable.

Al director de la DACYTI, el maestro Oscar Alberto González González, un agradecimiento particular pues ha sabido dar un seguimiento particular a las actividades realizadas durante mi estancia como estudiantes de posgrado, y es para mí un ejemplo de constancia y optimismo.

Han sido muchos los profesores de la UJAT, DACYTI y DACBiol que indirectamente han puesto un granito de arena en este trabajo como para mencionarlos a cada uno de ellos, pero a todos por igual quiero ofrecerles mi más sincero agradecimiento por haberme permitido aprender de ellos.

Agradezco de igual manera a mi compañero y amigo Juan Pablo Quiñonez (QEPD) por ser una persona extraordinario, ejemplo de vida, superación y lucha.

A mi comité revisor, porque gracias a ellos y a sus aportaciones fue posible mejorar en gran medida el desarrollo y entendimiento de este proyecto de investigación que hoy culmina.

Resumen

La vaginosis bacteriana es una enfermedad común que aqueja a mujeres en edad de reproducción. Puede causar graves problemas de salud, sobre todo durante el embarazo. En su gran mayoría las personas infectadas no presentan síntomas detectables a simple vista, lo que dificulta su diagnóstico y, por tanto, su tratamiento.

En este trabajo de investigación se desarrollan diversos experimentos con métodos de aprendizaje automático, propios del área de la inteligencia artificial, para la identificación de los atributos más relevantes asociados al diagnóstico de la vaginosis bacteriana. Aunado a ello, se implementan diversas fases experimentales para hallar el modelo predictivo óptimo para el diagnóstico de la enfermedad. El propósito es recabar información para el desarrollo de técnicas mejoradas de detección de la vaginosis bacteriana con el mínimo de información posible.

Entre los hallazgos, se presentan dos rankings generales fundamentados en múltiples rankings individuales. Las corridas de métodos selectores de atributos son la base para la creación de dichos rankings. En ellos se muestran los atributos identificados como relevantes y se contrastan con los predictores, es decir, las características que aportan más valor predictivo al diagnóstico de la VB mencionados en las lecturas especializadas. Se destaca que, mediante los métodos implementados en este proyecto se señalan como relevantes otros predictores que los métodos tradicionales del área médica no indica.

Publicaciones

1. Jesús F. Pérez-Gómez, Juana Canul-Reich, José Hernández-Torruco y Betania Hernández-Ocaña. *Predictor Selection for Bacterial Vaginosis Diagnosis Using Decision Tree and Relief Algorithms.* Applied sciences. MDPI. Journal volume 10, issue 9, 3291. Mayo 2020.
2. Jesús Francisco Pérez-Gómez, Juana Canul-Reich y Erick De La Cruz-Hernandez. *Combinación de Rankings como Método para la Identificación de Biomarcadores de Vaginosis Bacteriana*. Research in computing science. Volume 149(8): 915-927. Agosto 2020.
3. Jesús Francisco Pérez-Gómez, Juana Canul-Reich y Rafael Rivera-López. *Support Vector Machine como método para la Identificación de Atributos Relevantes de Vaginosis Bacteriana*. Libro electrónico Futuro Digital: Avances y paradigmas tecnológicos. Pp. 394-399. Diciembre 2020.
4. Jesus Francisco Pérez-Gómez, Juana Canul-Reich y Erick De La Cruz-Hernandez. *An Enhanced Method for Diagnosis of Bacterial Vaginosis based on Support Vector Machines with Linear Kernel*. International Journal of Combinatorial Optimization Problems and Informatics. Septiembre 2021.

Índice general

Índice de figuras

Índice de tablas

Abreviaturas

VB **V**aginosis **B**acteriana
DT **D**ecision **T**rees -Árboles de Decisión-
ML **M**achine **L**earning -Aprendizaje Automático-
FSA **F**eature **S**election **A**lgorithms -Algoritmos de selección de atributos-
BV **B**acterial **V**aginosis -Vaginosis Bacteriana-
FP **F**alse **P**ositive
FN **F**alse **N**egative
SVM **S**upport **V**ector **M**achine -Maquinas de Vector Soporte-
TP **T**rue **P**ositive
TN **T**rue **N**egative
RF **R**andom **F**orests
RL **R**egresión **L**ogística

Capítulo 1. Introducción

1.1 Antecedentes

La microbiota es considerado como el conjunto de todos los microorganismos que subsisten en un ambiente. En el cuerpo humano podemos llegar a albergar hasta 100 billones de éstos y agruparse en un promedio de 1,000 especies diferentes [24].

La relación que existe entre el microbiota y el ser humano inicia desde antes de nacer. Hasta hace pocos años se pensaba que los humanos nacíamos estériles y que la colonización microbiana iniciaba inmediatamente al concebirse, la cual era altamente influenciada por el tipo de concepción [85].

Un ejemplo de esta relación se puede encontrar en la investigación realizada por Collado y otros [22], donde estudiaron el microbioma presente en heces maternales, placenta, fluido amniótico, calostro y meconio colectados de mujeres sometidas a parto por cesárea y sus bebés, en el cual, no solamente encontraron comunidades bacterianas en todas las muestras analizadas, sino que también se adjudicó que la placenta y el fluido amniótico comparten características comunes, caracterizadas por una muy baja diversidad y altos niveles de proteo bacterias comparadas con las otras muestras.

Normalmente, el ambiente vaginal sano alberga una gran diversidad de bacterias y microorganismos dentro de su hábitat. Si el equilibrio entre los diferentes tipos de bacterias que se encuentran en dicho ambiente es afectado (disbiosis), se pueden llegar a producir condiciones idóneas para la aparición y desarrollo de infecciones como la vaginosis bacteriana (VB), lo que impacta el estado de salud de las mujeres, el feto y los infantes recién nacidos.

Un ejemplo muy claro de las complicaciones que surgen de los cambios en la microbiota del tracto cérvico uterino es la aparición del Virus del Papiloma Humano (VPH), que puede volverse una amenaza persistente hasta el grado de convertirse en cáncer [33].

Prueba de ello es el estudio realizado por Atashili, Poole, Ndumbe, Adimora y Smith [3] el cual confirma que la vaginosis bacteriana está fuertemente asociada con el riesgo de contraer el Virus de Inmunodeficiencia Humana (VIH) al notar que la reducción de *lactobacillus* productores del peróxido de hidrógeno a su vez reducen las defensas vaginales contra microorganismos. Además, un alto índice de pH vaginal derivado de una infección incrementa la adherencia y la supervivencia de esta enfermedad. Éstos y otros cambios, combinados con las dificultades de erradicar exitosamente la vaginosis bacteriana, explican con certeza el aumento del riesgo observado en la mayoría de los estudios epidemiológicos.

A la fecha, cerca de 100 tipos de virus del VPH han sido identificados, de los cuales alrededor de 40 afectan el tracto genital, primordialmente trasmitidos a través del contacto sexual. Algunos de los factores claves responsables de la persistencia del VPH en pacientes incluyen la inmunodeficiencia, la edad, el fumar, anticonceptivos orales e infecciones como la *Chlamydia trachomatis* [105].

Basándose en datos como los anteriores, es claro que el microbiota humano se encuentra altamente asociado con funciones fisiológicas requeridas por el humano para mantener la salud.

La llegada de nuevas tecnologías concernientes a la adquisición, tratamiento, análisis y presentación de datos, el rápido desarrollo y la reducción en costos de la secuenciación masiva han dado pauta a una nueva perspectiva de estudio de microorganismos. La acelerada obtención y manejo de la variada cantidad de microorganismos en el ser

humano hacen de su estudio una tarea ardua, pues las posibles interacciones entre esas bacterias son demasiado extensas.

Además, el ruido en los datos obtenidos puede entorpecer la relación entre esas comunidades bacterianas. Esas dificultades son análogas a un problema encarado por investigadores, donde hay muchas interacciones genéticas posibles que pueden estar ligadas a muchas enfermedades conocidas y por conocer [5].

Ligada a la gran cantidad de datos obtenidos en el área de la biología y la necesidad del tratamiento con herramientas computacionales, surge la bioinformática como herramienta indispensable para investigaciones biológicas. Con la ayuda de ésta área de la informática podemos, entre otras cosas, predecir información genética, realizar análisis forenses, implementar algoritmos de comparación de secuencias, alineamiento de secuencias, contemplar interacciones intermoleculares, simulación molecular, análisis filogenéticos, análisis de relación a genes específicos y sus enfermedades, correlación de patrones y expresión genética, entre otras muchas cosas [65].

En la bioinformática pueden mezclarse múltiples algoritmos para la extracción, tratamiento, transformación, análisis, clasificación y presentación de datos según el enfoque del estudio. Sin embargo, en los últimos años el uso de modelos de aprendizaje automático -también conocido como *machine learning*- ha dado buenos resultados. *Machine learning* es un campo de las ciencias computacionales que usa algoritmos inteligentes para realizar predicciones basadas en los datos, y tiene un gran número de aplicaciones dentro del enfoque de la bioinformática. Permite, entre otras cosas, aplicar teorías matemáticas y computacionales para entender y procesar datos biológicos [19].

Uno de los campos en el área de la inteligencia artificial que brinda grandes resultados en la fase de preprocesamiento es el denominado método de selección de atributos. Es un campo de investigación nacido en los años setenta [30] que permite la eliminación

de características (atributos) irrelevantes y redundantes, proporcionando una mejor representación de la información original, lo que a su vez logra un costo computacional más bajo y una mejor interpretación por el algoritmo de aprendizaje [62].

De acuerdo con [37], la selección de atributos se enfoca en cuatro vertientes principales; facilita la visualización de datos y el entendimiento de los mismos, reduce los requerimientos de medición y almacenaje de los datos, reduce los tiempos de utilización y entrenamiento y define el curso de dimensionalidad para mejorar el rendimiento en la predicción.

1.2 Planteamiento del problema

La Vaginosis Bacteriana es una enfermedad relacionada con los cambios en el microbiota vaginal, y pueden o no presentarse síntomas que permitan conocer si una persona contiene esta infección. Según Jafarnejad y demás [50], esta infección es asintomática en un 50% a 75% de los casos, y en aquellos que pudieran manifestarse con efectos casi imperceptibles como el olor y el cambio de color. Este hecho complica el proceso de diagnóstico de la enfermedad.

Existen varios métodos para determinar la existencia de infección vaginal, y entre las más conocidas están los procedimientos clínicos denominados *Amsel criteria* y *Nugent scoring* que permiten diagnosticar la VB mediante la toma de muestra de fluido vaginal. El primero consiste en la observación de al menos tres de cuatro criterios; la coloración grisácea de la secreción vaginal, los niveles de pH por encima de los 4.5, la percepción de mal olor al agregar a la muestra hidróxido de potasio (KOH) o la presencia de células epiteliales alrededor de la vagina. El segundo método consiste en el sometimiento de la muestra de líquido vaginal a una tinción Gram, o coloración Gram; si después de la mezcla de varios líquidos de laboratorio el fluido final resulta en color morado, significa que la muestra contiene bacterias gram-positivas (alguna de las variedades de *lactobacilli*, *gardenerella vaginalis*, *prevotella* y/o *mobiluncus*), de lo contrario el

colorarse en rosa, rojo o grosella se considera que la muestra contiene bacterias Gram-negativas [50].

Un tercer método para la detección de la vaginosis bacteriana que es menos frecuente que las anteriores es producido por una reacción en cadena de polimerasa en tiempo real (PCR, por sus siglas en inglés) [46]. Este procedimiento valida la relación de *lactobacillus crispatus, iners, gasserí* y *jensenii* entre microorganismos anaerobios como *gardnerella vaginalis, atopobium vaginae, megasphaera* **phylotype.**

Los esfuerzos para caracterizar la vaginosis bacteriana usando métodos epidemiológicos, microscópicos, cultivo microbiológico y basados en secuencias han fallado al intentar revelar una etiología que sea consistentemente documentada en todas las mujeres afectadas [70]. Además, algunos de los procedimientos establecidos toman largos periodos de tiempo para analizar las muestras vaginales, sin olvidar que algunos de ellos son invasivos.

Un enfoque alternativo que pretende entender la etiología de la VB es el estudio de los microorganismos asociados a la enfermedad. Para ello se emplean análisis de biomarcadores -predictores-para detectar enfermedades o los procesos de las mismas. Este enfoque conlleva al estudio de los atributos del conjunto de datos y su relación con la vaginosis bacteriana. Por esta razón, este trabajo se aborda como un problema de selección de atributos.

Bases de datos actuales pueden contar con decenas e incluso cientos de miles de atributos con un alto grado de información tanto irrelevante como redundante. Esta gran cantidad de datos causa serios problemas a muchos algoritmos de minería de datos en términos de escalabilidad y rendimiento [66].

Para cualquier técnica de recolección de datos, se sabe por ende que la inversión de tiempo para realizarlo es uno de los factores clave en el diagnóstico de una

enfermedad. Optimizar los procesos para enfocarse en aquellos datos que realmente ofrecen información relevante debe ser esencial en todo proceso de análisis de la información. En enfermedades como la vaginosis bacteriana es tiempo es un elemento crucial para su tratamiento; por tanto, requerir pocos datos para diagnosticar la BV es útil para los médicos. De esto, surge la necesidad de identificar aquellos atributos del conjunto de datos que contengan mayor información acerca de la enfermedad para reducir el número de características a analizar para su diagnóstico. A estos atributos con mayor información de la enfermedad se les denomina atributos relevantes.

De acuerdo con Rubido [86]:

- Un atributo es fuertemente relevante si al eliminarlo del conjunto de datos afecta la precisión del clasificador, pues aporta información que ningún otro tiene, por lo que son atributos necesarios en el subconjunto óptimo.
- Un atributo es débilmente relevante si no es fuertemente relevante, pero bajo ciertas condiciones aporta información nueva, no siempre es necesario pues su información puede ser suministrada por un conjunto de atributos.

Generalmente para formular un diagnóstico, los médicos realizan preguntas relacionadas a los síntomas del paciente. A partir de este pequeño conjunto de datos, el medico forma un diagnóstico diferencial y decide qué características obtener (preguntas, exámenes, pruebas de laboratorio, historial, estudios de imágenes, etc.) para descartar diagnósticos en el conjunto de diagnósticos diferenciales. Con razonamiento hipotético-deductivo, las características más útiles son identificadas, de modo que cuando la probabilidad de una de los diagnósticos alcanza un nivel de aceptabilidad, el proceso es detenido, y el diagnóstico es aceptado. Por tanto, es posible obtener un nivel aceptable de certeza de el diagnostico con solo pocas características sin tener que procesar el conjunto completo de ellas [60]. De esta misma manera funcionan los algoritmos de selección de atributos, ya que mediante estos

métodos se identifican aquellas características con mayor relevancia para la detección de la enfermedad, lo que a su vez permitiría disminuir el número de atributos necesarios para su detección. La intención de implementar algoritmos de selección de atributos en un conjunto de datos como el de la vaginosis bacteriana es identificar los atributos más relevantes de esta afección -conocidos como biomarcadores en el área de la biología- y con base en ello determinar los más indispensables para la VB.

Cabe recalcar que en los artículos que se han presentado para la generación del estado del arte y los que se conocen de otros temas en relación a la bioinformática para el diagnóstico de VB donde el objetivo está planteado con base en el estudio y análisis de conjunto de datos públicos, se ha demostrado la eficacia de modelos con datos obtenidos y generados de manera colaborativa (gratuitos y disponibles en web), pero no se ha mostrado un proyecto en el cual el rendimiento y eficacia de los algoritmos de clasificación utilizados hayan sido puestos a prueba con datos reales y clínicos. A la aportación de datos clínicos reales y la ratificación por parte de un experto en el área de la biomedicina a los resultados de los algoritmos se le denomina validación clínica, lo que al proyecto le aporta mayor veracidad y confianza.

1.3 Línea de investigación

Los métodos, métricas y técnicas implementadas en este proyecto de investigación se centran en la recopilación, tratamiento y análisis de la información relacionada a la vaginosis bacteriana.

Por tanto, este trabajo se acopla a la línea de investigación en las ciencias de datos e inteligencia artificial. La finalidad, es aplicar las bases de la inteligencia artificial, y particularmente, el aprendizaje máquina, al descubrimiento de información subyacente entorno a la vaginosis bacteriana.

1.4 Preguntas de investigación

Para determinar la finalidad de esta investigación, se plantean las siguientes preguntas.

¿Cuáles son los atributos más relevantes de la vaginosis bacteriana identificados por métodos de selección de atributos?

¿La utilización de subconjuntos reducidos de atributos de la vaginosis bacteriana permiten mejorar el rendimiento de los modelos clasificadores?

¿Cuáles son los predictores imprescindibles entre los datos que permiten una óptima detección de la vaginosis bacteriana?

¿Cuántos predictores son indispensables para un desempeño óptimo de los modelos clasificadores?

1.5 Alcances de la investigación

Este proyecto de investigación es el primer intento por explorar una base de datos genuina y real respecto a la vaginosis bacteriana con métodos y herramientas propias del área del aprendizaje automático. La finalidad general del proyecto es recopilar información relevante que permita obtener un avance científico en el entendimiento de esta enfermedad. De manera particular, este proyecto gira en torno a la implementación de métodos de selección de atributos y métodos de clasificación, y se centra en los siguientes ejes:

- Exploración de diversos métodos de selección de atributos para identificar los biomarcadores más relevantes de la vaginosis bacteriana.
- Determinación del número óptimo de atributos que permitan obtener el más alto nivel de desempeño de clasificación con el menor costo computacional posible.
- Identificación de las características optimas de clasificación para la detección de la vaginosis bacteriana mediante los métodos propuestos.

- Determinación de una óptima combinación entre algoritmos selectores de atributos y algoritmos clasificadores para el diagnóstico confiable de la vaginosis bacteriana.

1.6 Objetivo general

Identificar los predictores relevantes de la vaginosis bacteriana de un conjunto de datos microbiológico y el modelo predictivo óptimo para su clasificación mediante la implementación de métodos de aprendizaje automático.

1.7 Objetivos específicos

- Identificar las variables predictoras más relevantes para la vaginosis bacteriana mediante la creación de rankings individuales con base en métodos de selección de atributos.
- Generar dos rankings generales de atributos mediante los rankings individuales.
- Implementar modelos de clasificación para el diagnóstico de la vaginosis bacteriana aplicando diversos métodos de clasificación.
- Implementar modelos clasificadores de la VB con la utilización de subconjuntos de datos creados a partir de los atributos identificados como más relevantes de la VB.
- Identificar el número óptimo de predictores de la vaginosis bacteriana que permitan el más alto desempeño posible de los modelos clasificadores.
- Validar la relevancia clínica de los resultados obtenidos.

1.8 Hipótesis

Mediante la identificación de atributos relevantes y la creación de modelos predictivos de la vaginosis bacteriana a través de métodos de aprendizaje automático se obtendrán modelos de clasificación con al menos 80% de precisión en la categorización de esta enfermedad.

1.9 Contribuciones

Si bien en este trabajo se implementan técnicas y métodos generalmente conocidos en el área del aprendizaje automático, este proyecto contribuye principalmente en el área de la microbiología en torno a la enfermedad de vaginosis bacteriana. Como parte del estudio microbiológico de los atributos de la VB, los resultados permitirán contrastar los atributos identificados como relevantes mediante ensayos clínicos con aquellos identificados como relevantes mediante técnicas de selección de atributos. La creación de modelos predictivos para la clasificación de la enfermedad contribuye al nivel de certeza en el diagnóstico de la VB. Estas aportaciones permitirían guiar a los interesados en la creación de un diagnóstico por computadora más rápido y altamente preciso para la detección de la enfermedad.

1.10 Organización

Este trabajo se distribuye de la siguiente manera. En el capítulo 2 se proporcionan los fundamentos y definiciones relacionadas a la vaginosis bacteriana y los métodos, técnicas y métricas del área de la inteligencia artificial y aprendizaje automático utilizadas en este proyecto. El capítulo 3 muestra algunas investigaciones y proyectos previamente desarrollados en el estudio de la enfermedad desde la perspectiva computacional. En el capítulo 4 se presenta el conjunto de datos de la vaginosis bacteriana utilizado en todo el proyecto y se detallan los experimentos realizados con los métodos de selección de atributo, así como los rankings generados con base en los resultados obtenidos. El proceso de construcción de los experimentos y resultados obtenidos de todos los escenarios desarrollados con los modelos predictivos se muestran en el capítulo 5. Finalmente, en el capítulo 6 se presentan las conclusiones y pretensiones para trabajos futuros.

Capítulo 2. Fundamentos

2.1 Definiciones del ámbito médico

2.1.1 Vaginosis Bacteriana

Es la más común de las infecciones vaginales conocidas que afecta a las mujeres en edad fértil, afecta a millones de mujeres en todo el mundo y causa problemas de salud graves [58]. Esta condición tiene un impacto decisivo en aspectos como la concepción, la capacidad de mantener un feto a término, el riesgo de adquisición de enfermedades de transmisión sexual (ETS) y la calidad de vida de las mujeres [94]. Aunque las causas que la desencadenan aún se desconocen, se sabe que es común en mujeres sexualmente activas. Afecta hasta el 29 % de todas las mujeres y está asociado con el riesgo de contraer enfermedades de transmisión sexual (ETS) y nacimientos prematuros [5]. La mayoría de las mujeres infectadas con esta enfermedad no muestran síntomas, pero es posible que se note una secreción vaginal blanca o gris y poco espesa, dolor, olor y tal vez ardor en el área vaginal.

La VB es una cambio complejo en la flora vaginal causado por una disminución en la prevalencia y concentración de H_2O_2 en el proceso de producción de lactobacilos [31] y un aumento en la prevalencia y concentración de microorganismos tales como *Gardnerella Vaginalis, Mycoplasma hominis;* bacilos anaeróbicos gram-negativos pertenecientes a los géneros *Prevotella, Porphyromonas* y Bacteroides*;* especies anaeróbicas como *Peptos-treptococcus, Mobiluncus, Mycoplasma, Corynebacterium, Enterococcus* [34,47,93].

2.1.2 Biomarcador

Un biomarcador (o marcador biológico) se refiere a una amplia categoría de señales médicas, es decir, indicaciones objetivas del estado médico observado externamente del paciente, que pueden medirse y reproducirse [96]. La Organización Mundial de la

Salud en coordinación con las Naciones Unidas definen un biomarcador como “Cualquier sustancia, estructura o proceso que puede ser medible en el cuerpo o su producto e influir o predecir la incidencia del resultado o la enfermedad” [106]. Para definir los microorganismos asociados, se utilizan análisis de biomarcadores, empleados para detectar enfermedades o sus procesos.

2.1.3 Microbiota

También conocida como comunidad microbiana, la comunidad de microorganismos que viven dentro y fuera de las cavidades humanas (fosas nasales, cara, dientes, aparato respiratorio, sistema intestinal, sistema urogenital, etc.) forman un ecosistema especifico que dependen heterogéneamente de la ubicación en el cuerpo y su genética [24].

2.1.4 Microbioma

El microbioma humano es el conjunto de genes de los organismos microscópicos (microorganismos) que se encuentran en nuestro cuerpo. Este conjunto de microorganismos se denomina microbiota, y principalmente está compuesto por bacterias, virus y hongos. En general, el microbioma bacteriano humano es predominante y tiene el mayor impacto sobre la salud. Más de cien mil billones de bacterias habitan en el organismo humano, siendo incluso 10 órdenes de magnitud más alto que el número de nuestras células. Aunque estos índices no se han logrado estimar con precisión, por el momento se sabe que millones de bacterias desempeñan un papel esencial en la regulación de numerosos procesos fisiológicos. Entre estos procesos cabe destacar la actividad de las enzimas digestivas, la síntesis de vitaminas del complejo B, la interacción con el sistema inmunológico, o la protección frente a organismos patógenos, entre otros [69].

2.2 Definiciones del ámbito computacional

2.2.1 Aprendizaje automático

Es una disciplina científica en el campo de la inteligencia artificial que se enfoca en desarrollar sistemas de aprendizaje para el reconocimiento de patrones completo entre millones de datos. Se trata de algoritmos que permiten, entre otras cosas, aprender de los datos y predecir comportamientos futuros [49].

Los algoritmos de aprendizaje automático se centran en realizar predicciones basada en generalidades de ejemplos previos y son ampliamente utilizados en la industria del negocio, ciencia, correo electrónico, gobierno, detección de fraude, negociación laboral, reconocimiento fácil y de iris, mensajes de correo electrónico, diagnóstico de enfermedades como el cáncer de seno, asma, demencia, y otras afecciones [4].

2.2.2 Bioinformática

La bioinformática se define generalmente como la aplicación de técnicas informáticas para ayudar a comprender y organizar la información asociada a los datos biológicos [65]. La bioinformática asiste en tres aspectos principales; Primero, permite la organización de los datos para que los investigadores accedan y aporten nuevos datos tanto como se encuentren; Segundo, permite el desarrollo de herramientas y fuentes para respaldar el análisis de los datos; Tercero, permite utilizar herramientas para analizar e interpretar los datos y su información de manera significativa.

2.2.3 Ranking de atributos

Muchos algoritmos de selección de atributos incluyen un proceso de ranking[1] de atributos, ya sea como mecanismo principal o auxiliar en la evaluación de los atributos

[1] El anglicismo “ranking” se refiere a una “clasificación” de elementos ordenados comúnmente de mayor a menor, útil para establecer criterios de valoración [79]. Para fines prácticos de esta investigación, nos referiremos a este término como ranking.

[37]. Uno de sus usos comunes es descubrir un conjunto de atributos líderes que más tarde puedan ser usados para crear un subconjunto de los datos [100]. La relevancia de los atributos es calculada con base en un criterio de ranqueo. Este criterio se define de acuerdo al método de selección de atributos implementado.

2.3 Métodos de clasificación

Son enfoques de aprendizaje supervisado donde un programa de computadora aprende de los datos de entrada y crea una clasificación para nuevas observaciones. Se le denomina supervisado por que se basan en la idea de que el método opera bajo supervisión al ser prevista del resultado real de cada uno de las muestras de entrenamiento [109]. En esta sección se detallan los diversos algoritmos y métodos de clasificación implementados en la fase experimental.

2.3.1 Máquina de vectores de soporte (SVM)

También es conocido como máquina de soporte vectorial, es un algoritmo de clasificación que crea un modelo que representa los puntos de muestra en el espacio de atributos y separa las clases tanto como sea posible [102]. El límite de decisión está representado por un hiperplano que crea un margen lo más grande posible de cada lado [39] (Véase la Figura 1). Cuando se evalúa una nueva instancia con un modelo creado con SVM, esta nueva instancia se coloca en una de las dos clases. Maximizar el margen entre las dos clases mejora el rendimiento del modelo predictivo [44].

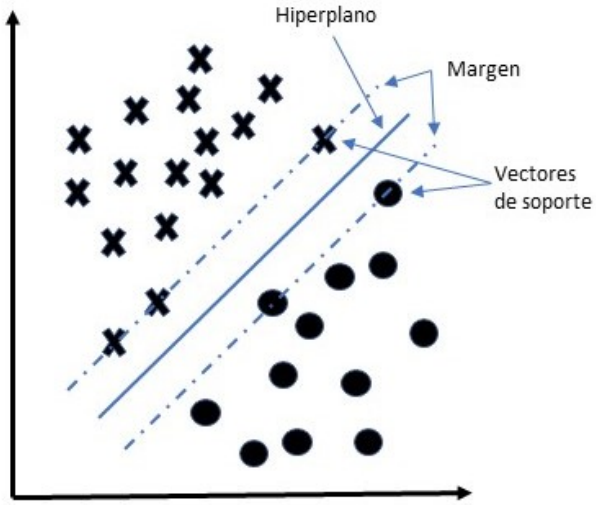

Figura 1. Representación gráfica de los vectores de soporte, margen e hiperplano.

SVM se basa en una función *kernel* que transforma los datos proporcionados de un espacio dado, también llamado “Espacio de entrada”, en un nuevo espacio dimensional llamado “Espacio de atributos”. En este espacio, los datos están separados por una superficie lineal llamada “Hiperplano” [18] (véase la Figura 2).

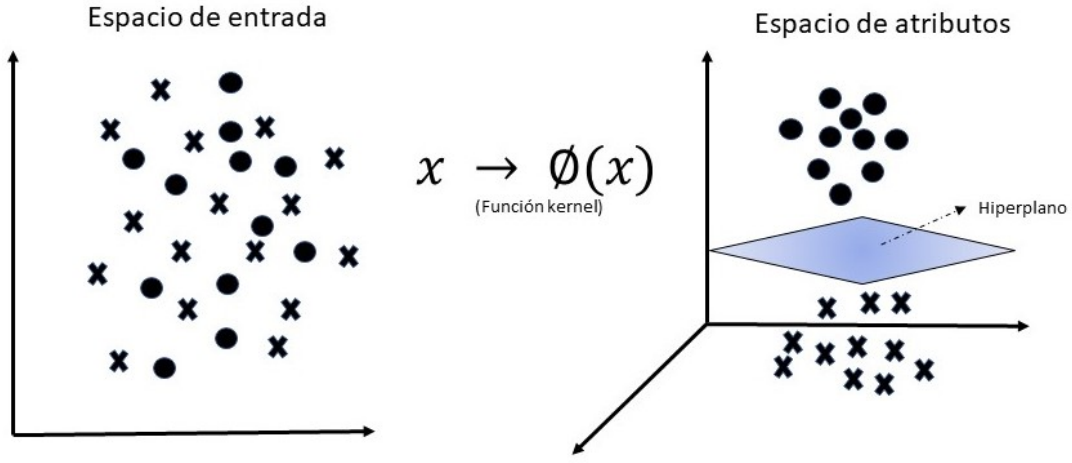

Figura 2. Los datos se transforman de un "Espacio de entrada" (izquierda) a un "Espacio de atributos" (derecha) mediante una funcion kernel.

Por tanto, donde los datos en un espacio de dos dimensiones son inseparables, ahora en un espacio de tres dimensiones es separable mediante un hiperplano.

Suponiendo que x_1 y x_2 son dos puntos de datos, $\emptyset$ es un mapeo y *K* denota la función *Kernel* dada por la Ecuación 2.1.

$$K\ (x_1, x_2) = \emptyset(x_1)^T\ \emptyset(x_2) \tag{2.1}$$

Un *kernel* toma dos argumentos, aplica un mapeo sobre ellos y devuelve el valor de sus productos. Un *Kernel* cuyo mapeo es idéntico, es decir, el espacio de entrada y el espacio de atributos son iguales, se llama *kernel* lineal. Una SVM que usa un *kernel* lineal, se le denomina SVM lineal [18]. Un *kernel* lineal se calcula usando la Ecuación 2.2.

$$K\ (x_1, x_2) = x_1^T x_2\ \Rightarrow\ \emptyset(x) = x \tag{2.2}$$

Una implementación del algoritmo SVM como método clasificador se proporciona en el paquete de *e1071* [17] en el lenguaje de programación R. En este trabajo, el algoritmo SVM se implementó con un *kernel* lineal y los parámetros por defecto. Esto es, Type= C-classification y cost=1.

Para fines prácticos en este trabajo, nos referiremos a este método de clasificación como *SVM,* por sus siglas en inglés.

2.3.2 Regresión Logística (RL)

Este método de clasificación se utiliza normalmente para predecir una respuesta binaria sobre uno o más predictores (atributos independientes o explicativas) utilizando un modelo lineal [71]. Se utiliza para calcular la probabilidad que ocurra una cierta clase, etiqueta o evento [42]. Las probabilidades de las posibles respuestas se modelan con la función logística representada por la Ecuación 2.3. Los valores de respuesta se codifican entre el 0 y el 1.

$$P(Y = 1 \mid X) = \frac{\exp\left(b_0 + \sum_{i=1}^{n} b_i x_i\right)}{1 + \exp\left(b_0 + \sum_{i=1}^{n} b_i x_i\right)} \tag{2.3}$$

Donde **P(Y=1|X)** es la probabilidad de que **Y** tome el valor 1 (presencia de la característica estudiada), **X** es un conjunto de *n* covariables **x1,⋯,xn** que forman parte del modelo, **b0** es la constante del modelo o término independiente y **bi** los coeficientes de las covariables. Para fines prácticos en este trabajo, nos referiremos a este método de clasificación como *RL,* por sus siglas en inglés.

2.3.3 Árboles de decisión (DT)

Este método de clasificación supervisado, también llamado a menudo reglas de decisión, surge de la idea estructural de un árbol formado por una raíz, nodos que separan las ramas, ramas y hojas. Un árbol de clasificación está conformado de nodos representados por círculos, y ramas representadas por los segmentos que conectan los nodos [32]. Un árbol comienza en la raíz, se extiende de arriba hacia abajo y se representa gráficamente de izquierda a derecha. El nodo inicial se le llama raíz, los nodos en los extremos se llaman hojas. Dos o más ramas pueden emerger de cada nodo interno [2]. La Figura 3 muestra un ejemplo de un árbol de decisiones.

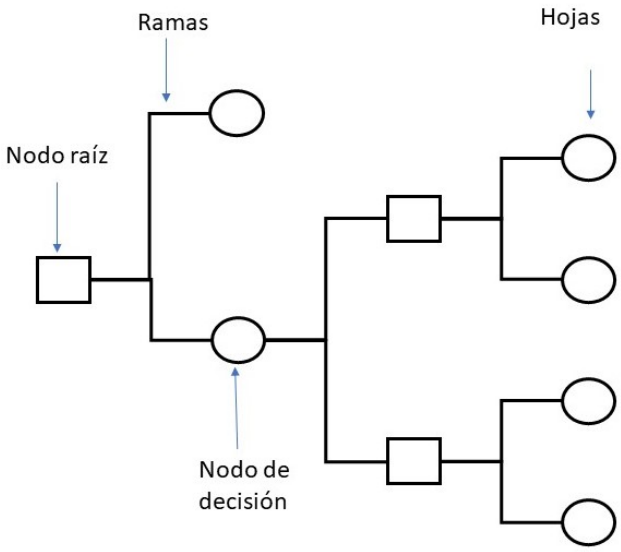

Figura 3. Representación gráfica de un árbol de decisión

En un árbol, la bondad de una división se cuantifica mediante una medida de impureza de los datos: una división es pura si, después de la división, todas las instancias de la elección de una rama pertenecen a la misma clase [32].

De acuerdo con [32], "Para un nodo m, Nm es el número de instancias de entrenamiento que alcanza el nodo m. Para el nodo raíz, esto es N. N_m^i pertenecen a las clases C_i, donde $\sum_i N_m^i = N_m$. Dado que una instancia alcanza el nodo m, la estimación de la probabilidad *p* de la clase C_i se define en la Ecuación 2.4".

$$\hat{P}(C_i \backslash x, m) == p_m^i = \frac{N_m^i}{N_m} \tag{2.4}$$

"El nodo m es puro si p_m^i para todo i son 0 o bien 1. Es 0 cuando ninguna de las dos instancias del nodo m son de clase C_i, y es 1 si todos estos casos son de C_i. Si la división es pura, no es necesario dividir más y se puede añadir un nodo hoja etiquetado con la clase para la cual p_m^i es 1".

La función para medir la impureza en los nodos es la entropía, y se calcula con la Ecuación 2.5.

$$I_m = -\sum_{i=1}^{K} p_m^i \, log \, p_m^i \tag{2.5}$$

Donde $0log0 == 0$. En el área de la teoría de la información, la entropía indica el número mínimo de bits necesarios para poder calcular la clase a la que pertenece una instancia.

Existen diferentes tipos de enfoque para la implementación de árboles de decisión. Específicamente para los experimentos planificados en esta investigación se utiliza el algoritmo denominado J48. Este algoritmo se basa en el algoritmo C4.5 diseñado por Ross Quinlan [77], y es un estándar de los algoritmos de árboles de decisión. El seudocódigo del algoritmo es mostrado en Algoritmo 1.

Algoritmo 1. Pseudocódigo del algoritmo J48 (Árboles de decisión).

```
1. Se crea un nodo raíz N;
2. Si (T pertenece a la misma categoría que C)
     {Nodo hoja = N;
     Marcar N como una clase C;
     Regresa N;
     }
3. Para i = 1 hasta n
     {Calcular la Ganancia_de_información (Ai);}
4. pa = atributo de prueba;
5. N.pa= atributo con la más alta Ganancia_de_información;
6. Si (N.pa== continuo)
     {Encontrar umbral;}
7. (Por cada T en la división de T)
8.       Si (T está vacío)
             {Hijo de N es un nodo de hoja;}
             De lo contrario
             {Hijo de N=dtree T;}
9. Calcula el rango de error de la clasificación del nodo N;
10. Regresa N
```

Para fines prácticos de este trabajo, nos referiremos a este método de clasificación simplemente como *DT,* por sus siglas en inglés.

2.3.4 Bosques aleatorios *(RF)*

Los bosques aleatorios son una combinación de árboles de decisión. La idea esencial es promediar muchos modelos basados en árboles, aproximadamente insesgados buscando reducir la varianza [44]. Este método se basa en un conjunto de árboles de decisión: “una muestra entra al árbol y es sometida a una serie de test binarios en cada nodo, llamados divisiones -*split*-, hasta llegar a una hoja en la que se encuentra la respuesta” [32]. Idealmente se forjó al utilizar la idea de dividir un problema complejo en un conjunto de problemas simples.

En la fase de entrenamiento, el método de bosques aleatorios optimiza los parámetros de las funciones de *split* a partir de las muestras de entrenamiento. Véase la Ecuación 2.6.

$$\theta_k{}^* = argmax_\theta \, j\epsilon\tau_j I_j \tag{2.6}$$

Donde $\theta_k{}^*$ es un vector aleatorio. Para ello, la función de ganancia de información es utilizada, la cual está dada por la ecuación 2.7.

$$I_j = H(j) - \sum i\epsilon 1{,}2 \frac{|S_j^i|}{|S_j|} H(S_j^i) \tag{2.7}$$

Donde S representa el conjunto de muestras que hay en el nodo a dividir, y S^i son los dos conjuntos generados a partir de la división [32]. La funciona mide la entropía del conjunto, y depende del tipo de problema que se plantea [12].

En Algoritmo 2 se presenta el pseudocódigo del algoritmo de bosques aleatorios.

Algoritmo 2. Pseudocódigo del algoritmo bosques aleatorios (*random forests*).
Para generar *c* clasificadores: 1. ***for*** *i*=1 ***to*** *c* ***do*** *2.* Muestra aleatoria del conjunto de entrenamiento *D* con reemplazo para producir *Di* 3. Crear un nodo raíz *Ni* que contenga *Di* 4. Llamar *Construccion_árbol (Ni)* 5. ***end for*** 6. *Construcción_arbol(N):* 7. ***if*** *N* contiene instancias de solo una clase ***then*** 8. ***return*** 9. ***else*** 10. Seleccionar aleatoriamente *x%* de los posibles atributos de división en *N* 11. Seleccionar los atributos *F* con la más alta ganancia de información para dividir 12. Crear *f* nodos hijos de *N*, *N1*, …, *Nf*, donde *F* tiene *f* posibles valores (*F1, …, Ff*) 13. for i=1 to f do 14. Colocar el contenido de *Ni* a *Di*, donde *Di* son las instancias en *N* que coinciden 15. *Fi* *16.* Llamar *Construcción_árbol(Ni)* 17. **end for**

Para fines prácticos de este trabajo, nos referiremos a este método de clasificación simplemente como *RF,* por sus siglas en inglés.

2.4 Métodos de selección de atributos

Estos son métodos de preprocesamiento computacional dedicados a buscar, seleccionar y/o discriminar una serie de variables que describen de manera eficiente la entrada de registros a la vez que se reducen los efectos de ruido o variables irrelevantes que no permiten buenos resultados de predicción. El principal objetivo de los métodos de selección de atributos es encontrar el mejor conjunto de atributos que permitan la construcción de modelos útiles para estudiar el fenómeno [89]. De acuerdo con [87], entre los aspectos más importantes en la implementación de métodos de selección de atributos se encuentran:

a) Evitar el sobreajuste y mejorar el rendimiento de los modelos
b) Proporcionar modelos más rápidos y a menor costo computacional
c) Obtener una visión más profunda de los procesos subyacentes que generaron los datos

Las técnicas de selección de atributos pueden ser organizadas en tres categorías dependiendo de cómo el método combina la búsqueda de atributos: métodos tipo filtro, métodos tipo envoltura y métodos embebidos -o incrustados-. Dichos tipos de métodos son descritos en las siguientes secciones.

2.4.1 Métodos de selección de atributos tipo filtro

Este tipo de métodos se basan en la información teórica de los atributos. Si un atributo no contiene información sobre la tarea realizada y a otras funciones ya seleccionadas, pueden omitirse del conjunto de datos con absoluta confianza [29]. Estos tipos de métodos seleccionan atributos basados en una puntuación o medida obtenida individualmente, es decir, evalúan cada atributo de forma aislada sin tener en cuenta la correlación entre ellos [63]. De esta manera es posible identificar aquellos atributos que son pobremente relevantes y aquellos que son fuertemente relevantes a la clase principal. La característica principal de los métodos con este enfoque es que se ignoran

por completo los efectos de los atributos en el comportamiento de un algoritmo clasificador [52].

1) Relief

Es un método tipo filtro multivariado para la selección de atributos. Su metodología para identificar atributos relevantes se basa en la distancia entre pares de instancias utilizando al implementar el algoritmo *nearest-neighbor* (vecinos más cercanos, por su traducción), el cual permite obtener una puntuación de relevancia para cada atributo del conjunto de datos [81].

El método *Relief* puede entenderse de la siguiente manera: para una instancia aleatoria (*Ri*), *relief* busca los dos vecinos más cercanos: uno de la misma clase (*nearest hit H*) y otro de clase distinta (*nearest miss M*). *Relief* restablece la estimación de calidad (*W[A]*) para todos los atributos *A* dependiendo de los valores para *Ri*, *M* y *H.* Si las instancias *Ri* y *H* tienen diferentes valores que el atributo *A*, entonces el atributo *A* separa dos instancias con la misma clase de modo que la estimación de calidad *W[A]* se decrementa. Por otro lado, si las instancias *Ri* y *M* tienen diferentes valores que el atributo *A*, entonces el atributo *A* separa dos instancias con diferentes valores de clase de modo que la estimación de calidad *W[A]* se incrementa. El proceso completo es repetido *m* veces, donde *m* es un parámetro definido por el usuario [81]. El seudocódigo del algoritmo *Relief* se muestra en Algoritmo 3.

Algoritmo 3. Pseudocódigo de *Relief*

Input: un vector de valores de atributos y el valor de clase para cada instancia de entrenamiento
Output: el vector *W* de estimaciones de las calidades de atributos
1. inicializar todos los pesos W[A: = 0.0
2. **for** *i: = 1* **to** *m* **do begin**
3. selecciona una instancia aleatoria *Ri*;
4. busca *nearest hit H* y *nearest miss M*;
5. **for** *A: = 1* **to** a **do**
6. *W[A]: = W[A] – diff(A, Ri, H)/m + diff(A, Ri, M)/m;*
7. **end**;

Donde la función *diff(A, I1, I2)* calcula la diferencia entre los valores de atributos *A* para dos instancias *I1* y *I2*. La función *diff* también se usa para calcular la distancia entre instancias para encontrar los vecinos más cercanos y se basa de la distancia Manhattan [81].

Una implementación del algoritmo *Relief* como método selector de atributos se proporciona con el paquete *FSelector* [83] en el lenguaje de programación R.

2) Chi cuadrada

Este método de tipo filtro evalúa la medida estadística mayormente conocida en el idioma inglés como *chi squared* para cada atributo con respecto a la clase principal [45]. Esta medida se basa en una prueba de hipótesis que compara la distribución observada de los datos con una distribución esperada de los datos. Este método de selección de atributos crea un ranking de atributos al considerar el chi cuadrado como el criterio de relevancia de los atributos en el conjunto de datos.

La prueba de chi cuadrado para un atributo f y la clase principal c se define en Ecuación 2.8.

$$x^2(f,c) = \frac{N[P(f,c)P(\bar{f},\bar{c}) - P(f,\bar{c})P(\bar{f},c)]^2}{(P(f)P(\bar{f})P(c)P(\bar{c}))} \tag{2.8}$$

Donde N es el numero de instancias en el conjunto de datos, $P(x,y)$ es la probabilidad conjunta de x y y, y $P(x)$ es la probabilidad marginal de x.

Para fines prácticos de este trabajo nos referiremos a este método selector de atributos como chi cuadrado. Una implementación del algoritmo chi cuadrado como método selector de atributos se proporciona con el paquete *snpStats* en el lenguaje de programación R.

3) Entropía

A pesar que el método *DT* es un algoritmo clasificador, también suele ser utilizado como selector de atributos, ya que, como parte de sus procedimientos éste utiliza un esquema de valoración para denotar un nivel de relevancia de cada atributo. Para ello, utiliza una medida llamada entropía. Esta medida se utiliza al construir un árbol de decisiones. Según Bramer [11], la entropía es una medida de la "incertidumbre" que contiene muchos datos de entrenamiento. La entropía se calcula al usar Ecuación 2.9.

$$E = -\sum_{i=1}^{K} p_i \log 2\, p_i. \qquad (2.9)$$

Si existen *K* clases en el conjunto de datos, se puede denotar la proporción de instancias con clasificación *i* como *pi* para *i=1* a *K*. el valor de *pi* es el número de ocurrencias de la clase *i* dividida por el número total de instancias, el cual es un número entre el 0 y 1, incluso ambos [11]. Cuanto menor sea la entropía, mayor será la ganancia de información de los atributos analizados [59].

Una implementación del algoritmo *decision tree* como método selector de atributos se proporciona con el paquete *caret* [55] en el lenguaje de programación R.

4) Ganancia de información

Fue descrito y presentado por Quinlan en 1986 [76] . Con este método se evalúa el valor de relevancia de un atributo midiendo la ganancia de información con respecto a la clase [15]. Mientras más alta sea la ganancia de información, mayor es el nivel de relevancia de los atributos. La ganancia de información se define como:

$$Ganancia\ de\ información(X|Y) = H(X) - H(X|Y), \qquad (2.9)$$

Donde $H(X)$ es la entropía de X, la cual se define como:

$$H(X) = -\sum_i P(x_i) log_2\big(P(x_i)\big), \tag{2.10}$$

Donde $P(x_i)$ es la probabilidad previa para todos los valores de X, y $H(X|Y)$ es la entropía de X dados los valores de la variable Y, la cual se define como:

$$H(X|Y) = -\sum_j P(y_j) \sum_i p(x_i|y_j) log_2(P(x_i|y_j)), \tag{2.11}$$

Donde $P\big(x_i|y_j\big)$ es la probabilidad posterior de x_i dado el valor de y_j.

En términos de clases y atributos, la ganancia de información como medida se expresa como:

$$InfoGain(Class|Attribute) = H(Class) - H(Class|Attribute). \tag{2.12}$$

Una implementación del método *information.gain* se encuentra disponible en el paquete *RWeka* [48] desarrollado para el lenguaje R.

5) Incertidumbre simétrica

El método es utilizado para medir la relevancia entre dos variables aleatorias [61] al calcular la incertidumbre simétrica con respecto a la clase [109]. Según Hall [40], se puede crear un modelo probabilístico de un atributo con valor nominal Y estimando las probabilidades individuales de los valores $y \in Y$ del conjunto de datos de entrenamiento. Si este modelo se usa para estimar el valor de Y para una nueva muestra, entonces la entropía del modelo (y por lo tanto del atributo) es el número de bits necesarios, en promedio, que se necesitarían para corregir la salida del modelo. La entropía es una medida de incertidumbre o imprevisibilidad en un sistema. La entropía de Y está dada por:

$$H(Y) = \sum_{y \in Y} p(y) \log_2(p(y)) \tag{2.13}$$

Si los valores observados de Y en el conjunto de datos de entrenamiento son particionados de acuerdo a los valores de un segundo atributo X, y la entropía de Y con respecto a las particiones inducidas por X es menor a la entropía de Y antes de la partición, entonces existe una relación entre las características Y y X. La ecuación 2.14 calcula la entropía de Y después de la observación de X.

$$H(Y|X) = -\sum_{x \epsilon X} p(x) \sum_{y \epsilon Y} p(y|x) \log_2\big(p(y|x)\big). \tag{2.14}$$

La cantidad por la cual la disminuye la entropía de Y refleja la información adicional sobre Y porporcionada por X, a la cual se le denomina ganancia de información. La ganancia de información se calcula con la ecuación 2.15.

$$Ganancia\ de\ información = H(Y) - H(Y|X) \tag{2.15}$$

$$= H(X) - H(X|Y)$$

$$= H(Y) + H(X) - H(X,Y)$$

La ganancia de información es una medida asimétrica -la cantidad de información ganada por Y después de la observación de X es igual a la cantidad de información ganada por X después de la observación de Y. La simetría es una propiedad deseable para una medida de inter correlación entre atributos [40]. La incertidumbre simétrica compensa el sesgo de la ganancia de información hacia los atributos con más valores al normalizarlos en el rango de 0 y 1 [74]. El coeficiente de la incertidumbre simétrica (CIS) se calcula con la Ecuación 2.16.

$$CIS = 2.0\left[\frac{gain}{H(Y)+H(X)}\right] \tag{2.16}$$

Con el método de incertidumbre simétrica es posible calcular la relevancia de los atributos utilizando el CIS. Esta medida cuantifica la intensidad relacional entre dos atributos, en este caso, la Inter correlación entre los atributos y la clase principal. Al tomar esta medida como criterio de relevancia, es posible la creación de ranking de atributos al considerar a los atributos con mayor CIS como los de mayor relevancia. Por el contrario, mientras menor sea su coeficiente de correlación, menor su relevancia dentro del conjunto de datos.

Una implementación del método *incertidumbre simétrica* para calcular la relevancia de los atributos se proporciona en el software *Weka* [108].

6) Longitud de descripción mínima (*MDL*)

Acrónimo de *minimum description lenght* (*MDL*) fue presentado por Kononenko en 1995 [53]. Este método trata el problema de selección de atributos como un problema de conexión; se construye un modelo con cada variable predictora y la etiqueta de clase [35].

En general, los fundamentos base de *MDL* establecen que la "mejor" teoría de los datos es la que minimiza la longitud o complejidad de la teoría y la longitud de los datos codificados en relación a la teoría [80]. Según Kononenko [53], tanto el emisor como el receptor tienen la descripción del número de atributos A, el número de valores posibles para cada atributo V, el número de clases C posibles y la descripción de los ejemplos de entrenamiento con respecto a los valores de atributo. Pero solo el remitente conoce la clasificación correcta de los ejemplos. Esta información debe transmitirse minimizando la longitud (bits) del mensaje. El remitente puede codificar explícitamente la clase para cada ejemplo de entrenamiento o seleccionar el "mejor" atributo y codificar, para cada

valor del atributo seleccionado, las clases de los ejemplos que tienen ese valor del atributo. Por tanto, se tiene un esquema de codificación para la distribución previa de las clases o tenemos un esquema de codificación separada para cada valor del atributo con la distribución posterior asociada. Para cada esquema de codificación, también debe transmitirse un decodificador.

Kononenko define una medida *MDL* de la calidad de los atributos mediante la Ecuación 2.17.

$$MDL = \frac{(Prior_{MDL} - Post\_MDL)}{n} \tag{2.17}$$

$$Prior\_MDL = \log_2 \binom{n}{n_1, \dots, n_C} + \log_2 \binom{n + C - 1}{C - 1} \tag{2.18}$$

$$Post\_MDL = \sum_j \log_2 \binom{n_j}{n_{ij}, \dots, n_{Cj}} + \sum_j \log_2 \binom{n_j + C - 1}{C - 1} \tag{2.19}$$

Donde n es el numero de instancias de entrenamiento, C es el numero de valores de clase, n_i es el número de instancias de entrenamiento para la clase C_i, n_j es el número de instancias de entrenamiento con el j-ésimo valor para el atributo dado, y n_{ij} es el número de instancias de entrenamiento de clase C_i que tiene el j-ésimo valor para los atributos proporcionados. $Prior\_MDL$ es la longitud de descripción de las etiquetas de clase previo al particionamiento entre los valores de un atributo. $Post\_MDL$ realiza el mismo calculo que $prior\_MDL$ por cada una de las particiones inducidas por un atributo y suma los resultados. Con las ecuaciones anteriores se codifican las etiquetas de clase con respecto al modelo codificado en el segundo término respectivo. El modelo *MDL* anterior es simplemente una distribución de probabilidad sobre las etiquetas de clase, es decir, cuantas instancias de cada clase están presentes; el modelo para $Post\_MDL$ es la distribución de la probabilidad de las etiquetas de clase en cada una de las particiones inducidas por el atributo dado.

Para obtener una medida que se encuentre entre 0 y 1, la ecuación 2.18 se puede normalizar al dividirla por $prior\_MDL/n$. Esto indica la fracción por la cual se reduce la longitud de la etiqueta de clase con la partición de los valores de un atributo. La ecuación 2.19 es una medida asimétrica; la compensación de roles de atributos y clases no suministran el mismo resultado. Para usar la medida asimétricamente para dos atributos, se puede calcular dos veces tratando cada atributo como una "clase" a su vez, y el promedio de los resultados.

Para fines prácticos, en este trabajo nos referiremos a esta medida *MDL* a la versión simétrica normalizada. Una implementación del método *MDL* para calcular la relevancia de los atributos se proporciona con el paquete *CORElearn* [82] del software R.

7) Puntuación de Fisher

Es un método supervisado de selección de atributos utilizado para la reducción de dimensionalidad de un conjunto de datos. Su objetivo es crear una puntuación para cada atributo de manera independiente bajo el criterio de Fisher para producir un subconjunto de características subóptimo al discriminar los de menor puntaje [36]. La idea general de este método es encontrar los atributos donde la distancia entre los puntos de diferentes clases sea lo más grande posible, mientras que la distancia entre los puntos de datos de la misma clase sea lo más pequeña posible [101]. El F-score, otro nombre con que se le conoce, proporciona una medida matemática de qué tan bien un atributo puede diferenciar entre diferentes clases [54].
La puntuación de Fisher se calcula utilizando la Ecuación 2.20.

$$F(x^j) = \frac{\sum_{k=1}^{c} n_k (\mu_k^j - \mu^j)^2}{{\sigma^j}^2} \tag{2.20}$$

Cuanto más alto sea la puntuación Fisher, mayor es el poder de discriminación. Mas información acerca de este método se encuentra en [36].

Para fines prácticos, en este trabajo nos referiremos a este método como *puntuación de fisher*. El paquete *PredPsych* [54] en el lenguaje de programación R proporciona una implementación del algoritmo.

8) Selección de atributos basada en correlaciones (*CFS*)

Mayormente conocido en inglés como *correlated-based feature selection (CFS)*, es un método selector de atributos de tipo filtro que permite evaluar dos aspectos de los predictores en un conjunto de datos: primero, la capacidad de predecir la clase a la cual pertenece; segundo, la correlación con otros predictores [40]. *CFS* ranquea los atributos y subconjunto de atributos de acuerdo a la función de evaluación heurística basada en correlación, representada por la Ecuación 2.21.

$$M_s - \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{2.21}$$

Donde M_s es el "mérito" heurístico de un subconjunto de atributos S que contiene k atributos, $\overline{r_{cf}}$ es la correlación media de clase/atributo $(f \in S)$, y $\overline{r_{ff}}$ es la correlación media de atributo/atributo. Mas detalles con respecto a este método selector de atributos se encuentra en [110].

Para fines prácticos, en este trabajo nos referiremos a este método simplemente como *CFS*. El paquete *FSelector* [84] en el lenguaje de programación R proporciona una implementación del algoritmo.

9) Correlación de Pearson

La correlación de Pearson (del inglés *Person's correlation*) es una medida estadística de correlación lineal entre dos variables [90]. Básicamente, esta medida indica la fuerza de la relación lineal entre dos atributos a y b [8] y se define mediante la Ecuación 2.22.

$$cc(a,b) = \frac{E(ab)}{\sigma_a \sigma_b} \tag{2.22}$$

Donde $E(ab)$ es la correlación cruzada entre a y b, y $\sigma_a^2 = E(a^2)$ y $\sigma_b^2 = E(b^2)$ son las varianzas de las señales a y b, respectivamente. Si el coeficiente de correlación $cc(a,b) = 0$, entonces entre a y b no existe una correlación. Cuanto más cercano este el valor de $cc(a,b)$ a 1, más fuerte será la correlación entre las dos variables. Si las dos variables son independientes, entonces $cc(a,b) = 0$. Información más detallada respecto a esta medida se encuentra en [8]. Por tanto, mediante el cálculo de esta medida es posible crear un ranking de atributos de la VB al considerar el coeficiente de correlación de cada atributo respecto a la clase principal. Mientras más cercano a 1 sea el coeficiente de correlación obtenido por los atributos, mayor será la relevancia para la VB.

Para fines prácticos, en este trabajo nos referiremos a este método como *correlación de Pearson.* Una implementación del método para el cálculo de *Pearson's correlation* se proporciona en el paquete *FSelector* [84] del lenguaje de programación R.

10) Correlación de Spearman

El coeficiente de correlación de Spearman (del inglés *Spearman's correlation*) mide la relación monótona de las variables en lugar de la asociación lineal tal cual lo realiza el método de Pearson. Por lo tanto, este método es más fiable con datos no lineales en comparación con la correlación de Pearson [107].

El método de correlación de Spearman computa la correlación entre el rango de x y el rango de y variables. Véase Ecuación 2.23.

$$rho = \frac{\sum(\acute{x} - m_{\acute{x}})(\acute{y} - m_{\acute{y}})}{\sqrt{\sum(\acute{x} - m_{\acute{x}})^2 \sum(\acute{y} - m_{\acute{y}})^2}} \tag{2.23}$$

Donde $\acute{x} = rank(x)$ y $\acute{y} = rank(y)$. Mas detalles del método se encuentra en [107].

Por tanto, mediante el cálculo de esta medida es posible crear un ranking de atributos de la VB al considerar el coeficiente de correlación de Spearman para cada atributo respecto a la clase principal. Mientras más cercano a 1 sea el coeficiente de correlación obtenido por los atributos, mayor será su relevancia para la VB.

Una implementación del método para el cálculo de *Spearman's correlation* se proporciona en el paquete *FSelector* [84] del lenguaje de programación R.

2.4.2 Métodos de selección de atributos tipo envoltura

Estos tipos de métodos involucran el desempeño del algoritmo de aprendizaje subyacente en el proceso de selección de atributos [16].

Los métodos de selección de atributos de tipo envoltura implementados en los experimentos se detallan a continuación.

11) Boruta

Es un método de selección de atributos tipo envoltura para la identificación de atributos relevantes. Este algoritmo tiene como base el algoritmo llamado bosques aleatorios. El nivel de importancia de los atributos en el conjunto de datos se obtiene con base en la perdida de precisión de clasificación debido a la permutación aleatoria de los valores de los atributos [57]. El algoritmo funciona de la siguiente manera.

1. Extiende el sistema de información agregando copias (sombras) de todos los atributos (el sistema de información se expande en al menos 5 atributos sombra, incluso si el número de atributos en el conjunto original es menor a 5).
2. Mezcla los atributos agregados para eliminar sus correlaciones con el atributo respuesta.

3. Corre el clasificador de bosques aleatorios sobre el sistema de información extendido y recopila los *Z-score* computados.
4. Encuentra la máxima *Z-score* entre los atributos sombra (MZSA), y asigna un resultado a cada atributo que obtenga mejor puntuación con MZSA.
5. Para cada atributo de importancia indeterminada, realiza una prueba de calidad de dos lados con el MZSA.
6. Considera los atributos que tienen una importancia significativamente menor que MZSA como “sin importancia” y los elimina permanentemente del sistema de información.
7. Considera los atributos que tienen una importancia significativamente mayor que MZSA como “importante”.
8. Remueve todos los atributos sombra.
9. Repite el procedimiento hasta que la importancia es asignada para todos los atributos, o hasta que el algoritmo alcance el límite establecido de ejecuciones del algoritmo clasificador.

Al final, el método *Boruta* devuelve una puntuación (medida de relevancia) para cada atributo en el conjunto de datos de acuerdo a su proceso de evaluación de atributos.

Una implementación del algoritmo *boruta* como método selector de atributos se proporciona con el paquete *Boruta* en el lenguaje de programación R.

12) Consistencia

También suele llamarse rango de consistencia (del inglés *consistency)* es una medida y un criterio de selección que, a diferencia de la mayoría, no busca maximizar la separabilidad de las clases, sino mantener el nivel de discriminación de los datos definidos por los atributos originales [25]. Con esta medida, la selección de atributos permite buscar el conjunto de atributos más pequeño posible que pueda distinguir entre una u otra clase tal cual como si se utilizara el conjunto de atributos completo.

La medida *consistencia* se aplica a la selección de atributos de la siguiente manera: dado un subconjunto de atributos candidatos S se calcula su "rango de inconsistencia" $I_R(S)$. Si $I_R(S) \leq \delta$, donde δ es un umbral de tasa de inconsistencia proporcionado por el usuario, el subconjunto S es consistente. Finalmente, una corrida del método proporciona como salida el subconjunto de atributos óptimo definido por la medida *Consistencia*.

Información más detallada respecto a este método como selector de atributos se puede hallar en [91]. Una implementación del algoritmo *Consistencia* como método selector de atributos se proporciona con el paquete *FSelector* [84] en el lenguaje de programación R.

13) Selección secuencial hacia adelante (*SFS*)

Del inglés *Sequential Fordward Selection* (*SFS*), es un método que identifica el "mejor" subconjunto de atributos basado en el número de ocasiones en la que cada uno de ellos se selecciona en el número de repeticiones dado [51]. En cada corrida del método, se evalúan los atributos con base en la tasa de clasificación correcta más alta respecto al método de clasificación implementado.

Este método de búsqueda ascendente, como también es conocido [67], comienza desde un conjunto de atributos vacío S y agrega gradualmente atributos seleccionados por una función de evaluación, que minimiza en error cuadrático medio (MSE, por sus siglas en inglés). En cada iteración, el atributo que se incluirá en el conjunto de atributos se selecciona entre los atributos restantes disponibles del conjunto de atributos, que no se han agregado al conjunto de atributos. Por tanto, el nuevo conjunto de atributos extendido debería producir una tasa de error de clasificación mínimo en comparación con la adición de cualquier otra característica. Mas detalles respecto a este método se encuentra en [51,67].

Una implementación del método para calcular la relevancia de los atributos se proporciona con el paquete *mlr* [9] del software R.

Para fines prácticos de este trabajo, nos referiremos al método como *SFS*.

14) Selección secuencial hacia atrás (*SBS*)

Del inglés *sequential backward selection* (*SBS*) es un algoritmo de selección de atributos que reduce el espacio de atributos en un subconjunto optimo con una latencia mínima en el rendimiento del clasificador y a la vez que disminuye el tiempo de ejecución del modelo [43]. Con este método, los atributos se eliminan secuencialmente del conjunto de atributos hasta que el nuevo subconjunto de atributos tenga suficientes características. Para calcular los atributos a eliminar, es necesario definir una función de criterio para minimizar [75]. El criterio se basa en la diferencia del rendimiento del clasificador antes y después de la eliminación de un atributo en específico.

El funcionamiento general del método se describe a continuación. Primero, la función de criterio es calculado para todos los atributos n en el conjunto de datos. Entonces, cada atributo es eliminado uno a la vez, la función de criterio es calculado para todos los subconjuntos con atributos $n-1$, y el "peor" atributo se descarta. Posteriormente, cada atributo del restante $n-1$ es eliminado uno a la vez, y el peor atributo se descarta para formar un nuevo subconjunto con atributos $n-2$. Este procedimiento continúa hasta que quede un numero predefinido de atributos. Mas detalles respecto al funcionamiento de este método de selección de atributos se puede consultar en [75].

Para fines prácticos de este trabajo, nos referiremos a este método como *SBS*. Una implementación de este método para calcular la relevancia de los atributos se proporciona con el paquete *mlr* [9] del software R.

15) Selección flotante secuencial hacia adelante (SFFS)

Del inglés *sequential forward floating selection (SFFS)* es básicamente una técnica de búsqueda ascendente que agrega atributos aplicando el método *SFS* -descrito con anterioridad- a partir del conjunto inicial atributos, seguido de una serie de exclusiones condicionales consecutivas del peor atributo en el conjunto de atributos recién creado, siempre que se pueda realizar una mejora adicional en los subconjuntos anteriores. *SFFS* comienza a partir de un modelo de atributos vacío, después de cada paso hacia adelante, *SFFS* retrocede hasta donde la función objetivo aumente. Mas detalles del algoritmo *SFFS* se encuentra en [75].

Una implementación del método *SFFS* para calcular la relevancia de los atributos se proporciona con el paquete *mlr* [9] del software R.

16) Selección secuencial flotante hacia atrás (SBFS)

Del inglés s*equential backward floating selection (SBFS),* es un procedimiento de búsqueda de “arriba hacia abajo” que excluye atributos mediante la aplicación del método *SBS* (descrito anteriormente) a partir del conjunto de atributos actual y seguido de una serie de inclusiones condicionales sucesivas del atributo más significativo de los atributos disponibles si se produce una mejora a los conjuntos de atributos previos. *SBFS* (Por sus siglas en inglés) comienza con un modelo de atributos completo, en cada paso el algoritmo elige el mejor modelo de todos los modelos con un atributo adicional y de todos los modelos con una característica menos. Mas detalles del algoritmo *SBFS* se halla en [75].

Una implementación del método *SBFS* para calcular la relevancia de los atributos se proporciona con el paquete *mlr* [9] del software R.

2.4.3 Métodos de selección de atributos tipo embebidos

Este tipo de métodos incluye una búsqueda y evaluación de un subconjunto de atributos al crear un modelo clasificador [38]. Los métodos embebidos tienden a ser más eficientes computacionalmente que los otros tipos, ya que integran de manera simultánea el modelado del clasificador con el proceso de selección de atributos. Esto puede hacerse, por ejemplo, optimizando una función objetivo de dos partes con un término de bondad y ajuste, o una penalización para un mayor número de características [100]. A continuación, se describen los métodos de selección de atributos de este tipo implementados en esta investigación.

17) Peso de los atributos

Mediante el proceso de creación del modelo clasificador con *SVM*, es posible determinar el nivel de relevancia de los atributos en el conjunto de datos mediante el peso de los atributos, mayormente conocido en el lenguaje inglés como *feature weights* [23][15]. Los pesos w de cada atributo se calculan mediante la Ecuación 2.24.

$$w = \sum_{i=1}^{nSV} \alpha_i y_i x_i \tag{2.24}$$

Donde nSV es el numero de vectores de soporte que son las únicas muestras de entrenamiento con valores alfa dintintos de cero, y_i son las etiquetas de clase (1/0) para el *n*-esimo vector soporte, a_i es un valor positivo real que indica la contribución al margen del *n*-ésimo vector soporte al modelo SVM, y x_i es el valor del atributo en el *n*-ésimo vector soporte [23].

Una implementación del método *SVM* para calcular la relevancia de los atributos mediante los pesos se proporciona con el paquete *e1071* [17] del software R.

18) Regresión logística

En la creación de un modelo de clasificación mediante una regresión logística es posible calcular el peso de los atributos utilizando el Coeficiente de Correlación (CC). Esta medida cuantifica la intensidad de relación entre dos atributos. Al tomar esta medida como criterio de relevancia, es posible la creación de ranking de atributos al considerar a los atributos con mayor CC como los de menor relevancia. Por el contrario, mientras menor sea su coeficiente de correlación, menor su relevancia dentro del conjunto de datos.

Una implementación del método RL para calcular la relevancia de los atributos mediante los pesos se proporciona con el paquete *caret* [55] del software R.

19) OneR

OneR, abreviatura de *One Rule* (Una regla, por su traducción) es un algoritmo de clasificación que genera un árbol de decisión de un sólo nivel y es capaz de inferir una clasificación mediante reglas de un conjunto de instancias [56]. El método crea una regla de clasificación para cada atributo en el conjunto de datos de entrenamiento, y selecciona la regla con el margen de error más bajo. A esta regla le denomina *'one rule'*.

A pesar que es un método un clasificador, mediante su procedimiento es posible calcular el nivel de relevancia de los atributos, por lo que también se considera un método de selección de atributos tipo embebido. Para ello, se determina la clase de mayor frecuencia de cada valor de atributo en la creación de la regla. La clase que aparece con más frecuencia en más ocasiones es la clase con mayor frecuencia para ese valor de atributo. Una regla es simplemente un conjunto de valores de atributos asociados a su clase mayoritaria [56]. Para cada atributo, el algoritmo crea una regla simple al determinar la clase mayoritaria de ese atributo. Luego, evalúa los niveles de precisión con cada una de las reglas y los atributos se ordenan de acuerdo a la calidad del modelo [73].

El seudocódigo implementado para el algoritmo de *OneR* se proporciona en el Algoritmo 4.

Algoritmo 4. *OneR*
1. Por cada predictor 2. Por cada valor de ese predictor, crear una regla: 3. Contar la frecuencia de aparición de cada valor de la clase objetivo 4. Encontrar la clase más frecuente 5. Hacer que la regla asigne esa clase a este valor del predictor 6. Calcular el error total de las reglas de cada atributo 7. Elegir el predictor con el error total más bajo.

OneR devuelve una lista que contiene la llamada a la función con los argumentos especificados, el nombre de los atributos y la clase objetivo, una lista de reglas, el número de instancias totales y correctamente clasificadas, y la tabla de contingencia del mejor predictor *versus* la clase objetivo.

Una implementación del algoritmo *OneR* como método selector de atributos se proporciona con el paquete *OneR* en el lenguaje de programación R.

20) LASSO

El método *Least Absolute Shrinkage and Selection Operator (LASSO)*, también conocido como regresión penalizada, fue formulado por Robert Tibshirani en 1994 [97]. Éste, permite reducir algunos coeficientes de los atributos a cero, los cuales pueden eliminarse del modelo sin impactar la capacidad para predecir el resultado de interés [98].

LASSO permite minimizar la suma de los errores cuadrados, con un límite superior para la suma de los valores absolutos de los parámetros del modelo. La formulación usada por Bühlmann y Van de Geer [13] es la siguiente:

La solución de LASSO se define por la solución del problema de optimización l_1 al minimizar:

$$\left(\frac{\|Y - X\beta\|_2^2}{n}\right) \tag{2.25}$$

Sujeto a:

$$\sum_{j=1}^{k} \|\beta\|_1 < t \tag{2.26}$$

Donde t es el limite superior para la suma de los coeficientes. Este problema de optimización es equivalente a la estimación de parámetros siguiente:

$$\hat{\beta}(\lambda) = \underset{\beta}{argmin}\left(\frac{\|Y - X\beta\|_2^2}{n} + \lambda\|\beta\|_1\right) \tag{2.27}$$

donde $\|Y - X\beta\|_2^2 = \sum_{i=0}^{n}(Y_i - (X\beta)_i)^2$, $\|\beta\|_1 = \sum_{j=1}^{k}|\beta_j|$ y $\lambda \geq 0$ es el parámetro que controla la fuerza de penalización, cuanto mayor es el valor de λ, mayor es la cantidad de reducción.

De acuerdo con la definición de [13], “La relación entre λ y el límite superior t es una relación inversa. De hecho, cuando t se vuelve infinito, el problema se convierte en mínimos cuadrados ordinarios, y λ se convierte en 0. Y viceversa, cuando t se vuelve 0, todos los coeficientes se reducen a 0 y λ tiende a infinito”.

Para la propiedad de LASSO como selector de atributos, algunos coeficientes se reducen a 0 al minimizar el problema de optimización, Por lo tanto, los atributos con coeficiente igual a 0 pueden ser excluidos del modelo. Al considerar como criterio de

relevancia los coeficientes de los atributos obtenidos mediante LASSO es posible crear un ranking de atributos.

para fines prácticos de esta investigación, nos referiremos a este método como *LASSO*. Una implementación del método *LASSO* para calcular la relevancia de los atributos se proporciona con el paquete *glmnet* [92] del software R.

21) Bosques aleatorios regularizado

El método *regularized random forests (RRF),* por cómo es mayormente conocido [27], aplica el marco de regularización del árbol al *RF* y puede seleccionar un subconjunto compacto de atributos.

Deng & Runger (2013) menciona que la ganancia de información, entendida como $Gain_R(X_i, v)$ es usada en *RRF* como:

$$Ganancia_R(X_i, v) = \begin{cases} \lambda \cdot Ganancia(X_i, v) & i \notin F \\ Ganancia(X_i, v) & i \in F \end{cases}$$

Donde F es el conjunto de índices de los atributos usados para la división en nodos previos y es un conjunto vacío en el nodo raíz del primer árbol. El coeficiente de penalización se denomina $\lambda \in (0,1)$. Cuando $i \notin F$, el coeficiente penaliza el *n*-ésimo atributo para la división del nodo v. Un λ menor conduce a una penalización mayor. *RRF* usa $Gain_R(X_i, v)$ en cada nodo, y agrega el índice de un nuevo atributo a F si el atributo agrega suficiente información predictiva a los atributos seleccionados.

Si el coeficiente de penalización $\lambda = 1$, *RRF* tiene la regularización mínima. Aun así, un nuevo atributo tiene que ser más informativo en un nodo que los atributos ya seleccionados para ingresar al subconjunto de atributos. El subconjunto de atributos seleccionados por *RRF* es llamado *least regularized subset* -subconjunto menos regularizado-, lo que indica la mínima regularización de *RRF.*

El método *RRF* devuelve un nivel de relevancia de los atributos de acuerdo a las métricas utilizadas con anterioridad. Una implementación del método *RRF* se proporciona con el paquete *RRF* [26] del software R.

2.5 Métricas de desempeño de los modelos predictivos

2.5.1 Matriz de confusión

La matriz de confusión -también llamada tabla de contingencia- proporciona los resultados de rendimiento de un modelo predictivo durante la fase de "entrenamiento y prueba". Esta información se explica mediante una tabla con filas y columnas. Las filas representan las clasificaciones correctas. Las columnas corresponden a las clasificaciones predichas [10]. En la Tabla 1 se muestra un ejemplo de matriz de confusión para un modelo de clasificación binaria, es decir, de dos clases.

Tabla 1. Muestra de una matriz de confusión para una clasificación binaria.

		Predicción	
		Positivos	Negativos
Observación	Positivos	VP	FN
	Negativos	FP	VN

Donde VP -Verdadero Positivo-, VN -Verdadero Negativo, FP -Falso Positivo-, FN -Falso Negativo-. Una matriz de confusión es el punto de partida para el cálculo de las medidas de rendimiento de un modelo predictivo [45].

2.5.2 Precisión

Esta medida de rendimiento es visto como el criterio más importante entre las medidas de rendimiento de un modelo predictivo [10]. La precisión predictiva es la proporción del conjunto de instancias de pruebas que se catalogan como correctamente clasificadas.

En una clasificación binaria, la precisión se calcula al dividir el número de instancias correctamente clasificadas entre el número total de instancias. El cálculo de la precisión se define mediante la Ecuación 2.28.

$$Precisión = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.28}$$

En algunos documentos, ésta ecuación suele ser utilizada para el cálculo de otra métrica similar denominada “Exactitud”, pero en este trabajo se utiliza para expresar la “Precisión”.

2.5.3 Precisión balanceada

El conjunto de datos de vaginosis bacteriana utilizado en este proyecto de investigación se encuentra desbalanceado, esto es, la cardinalidad de las clases está muy alejadas. En otras palabras, el número de instancias entre las clases son remarcadamente diferentes. La precisión balanceada -también conocida como precisión ponderada- es el promedio de las precisiones obtenidas en todas las clases. Esta medida de rendimiento se calcula mediante la Ecuación 2.29.

$$Precisión\ balanceada = \frac{\left(\frac{VP}{VP + FN} + \frac{VN}{FP + VN}\right)}{2} \tag{2.29}$$

2.5.4 Sensibilidad

De acuerdo con Bramer [10], esta medida de rendimiento de modelos predictivos representa la proporción de instancias de prueba que son correctamente clasificadas como positivas. En el área de la medicina, esto es interpretado como el nivel de confianza de una prueba en generar correctamente un resultado positivo. Esta métrica se calcula mediante la Ecuación 2.30.

$$Sensibilidad = \frac{VP}{VP + FN} \tag{2.30}$$

2.5.5 Especificidad

Esta medida se considera como la proporción de instancias negativas que son clasificadas como correctas [10]. En el área de la medicina, este se interpreta como el nivel de confianza de una prueba en generar correctamente un resultado negativo. Por tanto, la especificidad se calcula mediante la Ecuación 2.31.

$$Especificidad = \frac{VN}{VN + FP} \tag{2.31}$$

2.5.6 Tiempo de clasificación

En este proyecto se considera el tiempo medido en microsegundos que le toma a un modelo de clasificación en ser entrenado y probado. Se calcula partir de que inicia la fase de entrenamiento hasta proporcionar los resultados de predicción. Para cuestiones prácticas de este proyecto, nos referiremos a esta medida como “Tiempo”.

2.5.7 Validación cruzada de *K*-pliegues

Es un método para obtener estimaciones confiables en conjuntos de datos pequeños y evitar el sobreajuste. Torgo (2010) describe este método de la siguiente manera: Obtenga *k* subconjuntos aleatorios y de igual tamaño de un conjunto de datos inicial. Para cada uno de los *k* subconjuntos se construye un modelo de clasificación utilizando los *k*-1 restantes para evaluar este modelo. El rendimiento del modelo se almacena y el proceso se repite para todos los subconjuntos restantes. Al final, existen *k* medidas de desempeño, todas obtenidas mediante la prueba de un modelo. Estos *k* rendimientos se

promedian, con el cual se obtiene un rendimiento general. En este proyecto, el valor de *k* que se utiliza para los experimentos es 10.

Capítulo 3. Revisión de literatura relacionada

3.1 Aprendizaje automático en el estudio de comunidades microbianas.

En su obra original, Halliday, McLellan, Amaral-Zettler, Sogin y Gast [41] utilizaron diferentes herramientas bioinformáticas para el análisis, estudio, comparación y visualización de datos genómicos extraídos de muestras de agua y arena para la comparación de las comunidades microbianas entre dos diferentes sitios en las playas de Avalon Bay, Australia, así como determinar la calidad del agua en los mismos puntos de interés. A través de las muestras físicas obtenidas y la extracción de los datos genómicos, se generaron 8050 secuencias de ADN que concierne todas las bacterias encontradas en dichas comunidades. Posteriormente utilizaron herramientas bioinformáticas tales como *PRIMER-E* (*Plymouth Routines in Multivariate-Ecological Research*) utilizada para el análisis de abundancia relativa, es un paquete estadístico para el análisis de datos a nivel de especies para comunidades ecológicas [21]; ANOSIM (*ANalysis of SIMilarities*), es una prueba no paramétrica (valor -*p*) y multivariada, comúnmente utilizada en el campo de la ecología que detecta las diferencias entre grupos de comunidades [20]; NMDS (*Non-metric Multi-Dimensional Scaling*) [14], utilizado para generar representaciones graficas de diferencias relativas en la composición de comunidades ecológicas; BIOENV, usado en las muestras para determinar cuál combinación de variables explica el mejor patrón en las secuencias de los datos de abundancia [104]; SIMPER, utilizado para identificar la secuencia especifica con la más alta contribución para el análisis de disimilitud observada entre las muestras [20]; VAMPIRE, es una colección integrada de herramientas que permite el análisis y visualización de datos para la estructura de población microbiana y su distribución.

Al final, el autor mencionó acerca de las diferencias y similitudes entre las dos muestras obtenidas inicialmente de los sitios analizados, así como la gran variedad de bacterias encontradas en cada ubicación.

Por otra parte, los autores Lu y demás [64] experimentaron con 20 ratones expuestos a arsénico dentro del agua durante cuatro semanas para determinar una posible relación entre el arsénico y el microbioma intestinal. A las cuatro semanas de exposición al arsénico, a los ratones se les aplicó la eutanasia y la necropsia para buscar patologías, inflamación, edemas, defectos en la piel, hiperplasia y displasia en múltiples regiones del cuerpo y del colon. Las heces fecales y la orina fueron colectadas durante la necropsia y se obtuvo los ADN aislados usando el Powersoil DNA Isolation Kit. El ADN fue amplificado usando cebadores universales en la región V4 del gen 16S rRNA de las bacterias. Las muestras individuales fueron secuenciadas usando Ilumina Miseq para generar las lecturas por pares. Utilizaron QIIME para filtrar la calidad, demultiplexar y analizar los archivos fastq. Se usó UCLUST para elegir las unidades taxonómicas operativas (OTU) con un umbral de 97% de similitud de secuencia.

Para el análisis estadístico fue utilizado el PCA (*Principal Component Analysis*) para examinar grupos metabólicos. El mapa de calor fue generado usando un algoritmo de agrupación jerárquica para visualizar las diferencias de metabolitos dentro del conjunto de datos. PCoA (*Principal Coordinate Analysis*) fue usado para comparar los perfiles del microbioma intestinal entre el control y tratamiento. La diferencia en la composición del microbioma intestinal fue evaluado más a fondo utilizando el software *Metastats*. La matriz de correlación entre los metabolitos relacionados a la microflora intestinal y las especies bacterianas del intestino fue realizada usando el coeficiente de correlación de Pearson.

El estudio de las secuencias obtenidas reveló que la exposición al arsénico alteró significantemente la composición de la micro flora intestinal de los ratones, mientras que

los experimentos metabólicos mostraron que un número determinado de metabolitos fueron perturbados substancialmente después de la exposición al químico. Adicionalmente, mediante el análisis de correlación se identificó que algunos grupos de bacterias se encuentran altamente correlacionadas con la micro flora intestinal alterada. Por tanto, mediante los resultados obtenidos se concluye que la exposición al arsénico no solamente perturba el microbioma intestinal, sino que también altera los perfiles de metabolitos, apoyando la hipótesis de que las perturbaciones del microbioma intestinal pueden servir como un nuevo mecanismo por el cual la exposición al arsénico provoca o exacerba las enfermedades humanas.

Otra investigación relacionada al estudio metagenómico con base en programas bioinformáticos es la que presentaron los autores Belda-Ferre y demás [7]. Mediante el proyecto obtuvieron 25 muestras de voluntarios que no cepillaron sus dientes las últimas 24 horas antes de tomar la muestra y posteriormente evaluado por un dentista. De dichas muestras, se extrajeron las secuencias de ADN usando MasterPure Complete DNA y RAN Purification Kit. El programa *Megablast* fue utilizado con el fin de identificar las secuencias humanas. Las lecturas metagenómicas se mapearon contra 1117 genomas de referencia secuenciados utilizando el algoritmo de alineación *Nucmer* y *Primer* v3.06. Se utilizó *RPSBlast* para alinear las lecturas de secuencias a los perfiles de proteínas.

Los análisis estadísticos realizados en el proyecto fueron hechos en *R*, mapas de calor de la composición taxonómica fueron generados el paquete *gplots* con frecuencia relativa por muestra, así como la distancia euclidiana y medianas normales. Se extrajeron secuencias de 16S rRNA de las lecturas de cada metagenoma por búsqueda de similitud usando *BLASTn*. La asignación filogenética de las secuencias se realizó utilizando el *RDP classifier*. La asignación taxonómica de todos los marcos de lectura abiertos se llevó a cabo con base en el algoritmo Lowest Common Ancestor (LCA) utilizando para eso el software *MEGAN*. Al final, todos los genomas completos y WGS

disponibles se recuperaron de la base de datos de microbiomas orales humanos y en NCBI todos los genomas bacterianos y de arqueas.

Finalmente, los autores mencionaron que una gran porción de ADN humano en las cavidades orales provee suficiente información genética para obtener información acerca de la microbiología de la caries, sugiriendo con base en el estudio realizado que la comunidad bacteriana es muy compleja, y que muestra que la placa dental de los individuos que nunca han sufrido caries puedan ser un reservorio genético de nuevos compuestos anti caries y probióticos.

Por su parte, McHardy [68] realizó un estudio basado en la mucosa intestinal, el cual es un sitio de compleja instrumentación de inmunología y metabolismo. Se considera que el estado inmunológico y funcional de la mucosa está influenciada por el microbiota. La composición microbiana está controlada, pero en ciertos individuos genéticos y ambientales resulta en manifestaciones clínicas como en enfermedades inflamatorias e inmunes. Realizaron un experimento con enfoque bioinformático para identificar la o las interrelaciones entre la composición microbiana y el intestino humano de 93 personas. Las muestras fueron obtenidas de las regiones cecum (ciego) y sigmoid. Los ADN fueron extraídos usando el kit *PowerSoil DNA Isolation*. Después de la preparación, la secuenciación de ADN fue realizada usando un *Illumina HiSeq 2000*. El control de calidad de las secuencias y la transformación a OTU's (*Operational Taxonomic Unit*) fueron realizados con *QIIME*. El número inicial de lecturas de secuencias obtenidas fue de 70,278,364, pero después de la limpieza de datos y el análisis de calidad se descartaron las secuencias de baja calidad, dejando 57,958,866 lecturas remanentes.

En este proyecto, todos los análisis fueron realizados utilizando el lenguaje estadístico *R*. La abundancia relativa fue realizada con una correlación de *Spearman* con el cálculo del *p-value* usando la función *cor.test*. Para el análisis de agrupamiento mediante mapa de calor se utilizó la función *heatMap.2* del paquete *gplots*. El

agrupamiento jerárquico de bacterias fue hecho con base en la métrica de distancia euclidiana, y el método completo de agrupamiento jerárquico por metabolitos. También utilizaron agrupamiento por *k-means* para evaluar la significancia en la predicción de cálculos fuertes, el cual ha demostrado ser una estrategia robusta para determinar el número de grupos óptimos para módulos de grupos jerárquicos. La fuerza de predicción fue realizado mediante la función *prediction.strength* del paquete *fpc* en R.

En conclusión, los datos revelan una interdependencia significante del metaboloma mucoso y el microbioma. La evidencia presentada sugiere que el microbioma y el metaboloma tienen influencia bidireccional, con una composición metabólica influencia por bacterias y la contribución de los metabolitos a la arquitectura de la comunidad microbiana. También sugiere que los metabolitos deberían ser más profundamente interrogados como un mediador directo de enfermedades asociadas a los microbios y que los metabolitos podrían ser un blanco directo para el monitoreo y manipulación terapéutica de comunidades microbianas y otras enfermedades asociadas al microbioma intestinal.

3.2 Aprendizaje automático en el estudio de la vaginosis bacteriana

Algunos de los trabajos de investigación realizados en torno al entendimiento de la vaginosis bacteriana desde la perspectiva del aprendizaje automático e inteligencia artificial se describen en el siguiente apartado.

Una investigación relacionada con el estudio de las infecciones vaginales como la vaginosis bacteriana es la que realizaron Ravel, Gajer, Abdo, Schneider, Koenig, McCulle y otros [78] en la cual caracterizaron el microbiota vaginal de 396 mujeres asintomáticas y sexualmente activas, representadas étnicamente en grupos; blancas, de color, asiáticas e hispanas. Para realizar el estudio genético se obtuvo una muestra del fluido vaginal para evaluar el microbiota utilizando el procedimiento Nugent para la detección de la vaginosis bacteriana, además de que se utilizó un procedimiento de

extracción del gen 16S rRNA en las regiones V1 y V2 para determinar la composición de especies y la estructura de las comunidades bacterianas residentes.

Después de un largo proceso químico-fármaco-biólogo para la obtención de la secuencia completa de ADN de las muestras obtenidas de las mujeres mediante PCR (Reacción en Cadena de Polimerasa, por su traducción al español), éstas fueron sometidas al clasificador Bayesiano con el Proyecto de Base de datos Ribosomal *(RDP,* por sus siglas en inglés) con la finalidad de obtener una asignación taxonómica en alguna de las especies registradas mediante un modelo denominado Markov.

El RDP es un clasificador Bayesiano que puede rápidamente, y con alto grado de efectividad, clasificar secuencias del gen 16S rRNA dentro del nuevo orden taxonómico. En el 98% de los casos donde la confidencia es más alta (mayor o igual a 95%) la tasa de efectividad y precisión llega a ser del 98%. Esta herramienta suele ser acompañada por la denominada *RDP Library Compare*, desarrollada para facilitar la comparación de comunidades bacterianas basadas en librerías de secuencias rRNA. Ambas son gratuitas y disponibles en línea Wang, Garrity, Tiedje y Cole [103].

Finalmente, con el estudio realizado se concluyó que las comunidades microbianas están agrupadas y dominadas por bacterias como *Lactobacillus Iners, L. crispatus, L. gasseri*. La proporción de cada comunidad bacteriana varió de acuerdo al grupo étnico de las mujeres muestreadas. Como dato extra obtenido mediante el proyecto, se encontró que el pH de las mujeres es diferente respecto al grupo étnico al cual pertenecen, predominando las hispanas con pH 5.0 y las de color con pH 4.7. Filo tipos (similitud y relación taxonómica entre un grupo de organismos) con abundancia relativa también fueron hallados en todas las comunidades, por lo que fueron asociados con los puntajes altos y bajos del criterio Nugent, los cuales son factor clave para el diagnóstico de VB.

A su vez, los autores Srinivasan y demás [95] secuenciaron la muestra de fluido vaginal de 242 mujeres con y sin vaginosis bacteriana mediante PCR (Reacción en Cadena de Polimerasa, por su traducción del inglés) con lo cual se obtuvo el gen 16S rRNA, y de igual manera, y como se había visto en otros estudios parecidos, ellos también utilizaron el procedimiento clínico denominado Criterios Amsel, y posteriormente confirmado con la tinción Gram para determinar VB.

Las secuencias fueron sometidas a un proceso de pre-procesamiento para asegurar la calidad de los datos y remoción de lecturas de baja calidad utilizando un paquete prediseñado llamado *microbiome* en *R-package*. Posteriormente los datos resultantes fueron utilizados para otras herramientas como *pplacer v1.1*, que busca la inserción optima de lecturas en un árbol filogenético de acuerdo a la máxima similitud o un criterio de probabilidad posterior bayesiano, lo que le permitió inferir la clasificación taxonómica.

La asociación entre la composición taxonómica de bacterias y el criterio Amsel fueron realizadas utilizando un modelo lineal penalizado denominado "*elasticnet*" el cual busca modelar un signo, por ejemplo, el pH vaginal, como una función lineal del logaritmo transformado en el conteo de bacterias de cada muestra. Además, se implementó el método de validación cruzada para valorar los resultados estadísticos.

Con base en los resultados se concluyó que cuando existe la ausencia de vaginosis bacteriana la comunidad de microorganismos es dominada por bacterias como *Lactobacillus crispatus* or *Lactobacillus iners*. Las bacterias *Leptotrichia amnionii* y *Eggerthella* fueron las únicas dos bacterias asociadas a VB, las cuales están altamente asociadas a los cuatro criterios Amsel. Los mismos resultados revelaron que la presencia de algunos subgrupos de bacterias asociadas a la vaginosis bacteriana sugieren una codependencia metabólica.

Por su parte, Beck y Foster [5] contribuyeron al estudio de la vaginosis bacteriana mediante su investigación. Ellos aplicaron tres algoritmos de aprendizaje automático

(*Genetic Programming*, *Random Forest* y *Regresión logística*) que han descubierto satisfactoriamente relaciones genéticas asociadas con enfermedades para descubrir interacciones con la VB. Su contribución se basa en la medición de la precisión de los tres algoritmos y en la clasificación de comunidades microbianas. Ellos utilizaron los bases de datos generados en los proyectos de Srinivasan y demás (2012) donde exponen los resultados del método clínico Nugent aplicado a 396 mujeres asintomáticas y del proyecto de Ravel y demás (2011) compuesto por datos de 220 mujeres a las cuales se les aplicó los criterios Amsel y puntuación Nugent. La evaluación de los algoritmos la hicieron aplicando el criterio de ROC (Curva del Operador Receptor, por su traducción del inglés) que muestra el rendimiento de los modelos al clasificar muestras de vaginosis bacteriana, tanto diagnostico positivo como negativo, además que permite comparar los márgenes de error de ambos modelos. Para la determinación de la precisión utilizaron la validación cruzada de diez pliegues; dividieron el dataset en diez partes, nueve para entrenamiento del modelo de clasificación y el restante para medir la precisión. Este procedimiento se repite intercambiando entre las diez partes del dataset.

Como resultados, los algoritmos *Random Forest* y *Lineal Regression* obtuvieron niveles de precisión entre 90% y 95% cuando se clasificaron datos de criterio Nugent para la detección de VB. Los mismos algoritmos resultaron relativamente bajos al clasificar datos clínicos obtenidos mediante el criterio Amsel, donde las tres técnicas de clasificación obtuvieron una precisión cerca del 80%.

Beck y Foster mostraron, finalmente, que el algoritmo *genetic programming* demostró ser más eficiente en la clasificación fenotípica basada en el criterio Nugent, pero no resultó muy confiable al clasificar los datos del criterio Amsel.

Éstos mismos autores (Beck & Foster, 2015) realizaron el mismo experimento, pero utilizando procedimientos que en el anterior estudio los consideraron innecesarios y

técnicas diferentes de clasificación de atributos. Aquí utilizaron los algoritmos de *random forest* y *regresión logística* para clasificar y modelar la relación entre el microbiota vaginal y la vaginosis bacteriana. El *dataset* utilizado fue el mismo que en el proyecto anterior. En este proyecto hicieron hincapié a la importancia en la utilización de todos los atributos del *dataset* para generar el rendimiento del clasificador (*accuracy*), y saber qué tan bien realizaban la clasificación de entradas nuevas los algoritmos utilizados. Con el fin de determinar cuántos atributos son necesarios para obtener un alto desempeño o precisión de los algoritmos, los atributos fueron agregados de manera secuencial en las pruebas realizadas a los algoritmos. La precisión de cada prueba para cada clasificador fue determinada utilizando las técnicas conocidas como *ROC* y *AUC* (por sus siglas en inglés).

A su vez los autores Baker, Agrawal, Foster, Beck, & Dozier (2014) se propusieron descubrir los atributos más significantes necesarios para diagnosticar la vaginosis bacteriana aplicando varios algoritmos de clasificación y selección de atributos. La selección de atributos es el proceso de elegir las más significantes características o atributos y la formación de subgrupos que serán los más valiosos para el análisis de predicción. Entre los algoritmos utilizados se encuentran *CfsSubsetEval, ClassifierSubsetEval, ConsistencySubsetEval, FilteredSubsetEval and WrapperSubsetEval,* todos propios del entorno *Weka.* Para realizar la evaluación de los resultados obtenidos y predecir la clase de cada instancia, utilizaron algoritmos de clasificación como *Bagging, RBFNetwork, J48, NaÏve Bayes, AdaBoostM1, RandomForest, LogitBoost, KStar (K*), FT*.

Los datos utilizados tienen un total 1601 instancias y 418 atributos, los cuales fueron obtenidos de 25 mujeres a lo largo de 10 semanas, y está compuesto por la combinación de resultados de un cuestionario, de análisis médicos y análisis clínicos como el estudio de criterios *Amsel*. El cuestionario incluyo preguntas acerca de la actividad sexual, uso de métodos anticonceptivos, uso del tabaco, entre otras cosas.

Los datos médicos fueron obtenidos mediante una muestra vaginal que fue sometida a secuenciación del gen 16S rRNA para su identificación taxonómica a nivel especie.

Los resultados obtenidos muestran que el algoritmo FT (*Functional trees*) es el mejor algoritmo para este caso, considerando su tiempo de ejecución, la reducción de atributos y el *recall*. Para los algoritmos de selección de atributos los algoritmos que tuvieron mayor reducción de atributos con 14 de ellos fueron *WrapperSubsetEval* con el método *BestFirst* y con el método *GreedyStepwise.*

Para trabajos futuros proponen realizar una limpieza del conjunto de datos antes de utilizarlos, remover las series de tiempos utilizados y realizar pruebas con los datos médicos y datos clínicos de manera separada.

En conclusión, los autores pudieron generar una tabla donde expresaban los quince atributos más importantes para cada clasificador con los que lograron una precisión muy alta, además pudieron confirmar que solamente un par de atributos son necesarios para generar modelos con un alto desempeño para la clasificación de la vaginosis bacteriana, siempre y cuando se determine como importantes (biomarcadores) mediante los algoritmos de selección de atributos.

3.3 Selección de atributos en el estudio de la vaginosis bacteriana

Han sido pocas las investigaciones que se han desarrollado en torno al estudio de biomarcadores en el área de la vaginosis bacteriana. Sin embargo, el impacto de dichos trabajos ha tenido repercusión positiva para el estudio de la afección.

Para dar un ejemplo, el trabajo de Yolanda Baker y demás [4] tuvo como propósito descubrir los atributos más significantes de la vaginosis bacteriana y la aplicación de algunos algoritmos de clasificación para el diagnóstico de la enfermedad. Veinte algoritmos de selección de atributos, en combinación con nueve algoritmos clasificadores se aplicaron utilizando la herramienta Weka. Medidas de rendimiento de

clasificación como la precisión, el *recall* y el número de atributos reducidos se registraron en los experimentos. Mediante los resultados experimentales se determinó que los algoritmos *Functional Trees* -una generalización de árboles multivariados- implementado como clasificador y *WrapperSubSetEval* usado como selector de atributos produjeron la mejor combinación.

En una de sus investigaciones, los autores Beck y Foster [5] implementaron tres métodos de aprendizaje máquina para estudiar las posibles interacciones microbianas asociadas con la vaginosis bacteriana. Entre ellos los algoritmos *genetic programming (GP), random forests (RF)* y *regresión logística (RL)*. Mostraron que dichos métodos pueden clasificar, con alta precisión a las mujeres según el estado de la VB de acuerdo al microbioma vaginal y algunos factores ambientales. Principalmente, los autores estaban interesados en dos aspectos de los modelos clasificadores, la precisión en el modelo de diagnóstico y el nivel de certeza respecto al número de atributos utilizados. Para los experimentos, se usaron dos conjuntos de datos microbiológicos diferentes para entrenar y evaluar los métodos, uno de Srinivasan y Fredericks [95] y el otro de Ravel y demás [78]. Entre los resultados, se mostró que modelos de RF y RL obtuvieron entre el 90% y el 95% de precisión al clasificar la VB. Las diferentes técnicas de clasificación variaron ampliamente en el tiempo computacional. *RL* y *RF* fueron relativamente rápidos, por lo general se completaban en menos de una hora. *GP* tardó varias horas. Mediante los métodos implementados, también se definieron niveles de relevancia para cada atributo en el conjunto de datos, lo que permitió identificar los predictores con mayor relevancia de la VB. En el caso de *GP*, el nivel de relevancia de los atributos se determinó mediante la “Pureza” en el nodo, una medida computada para cada atributo que calcula la separación entre muestras BV+ y BV- para cada árbol de decisión, que al final se promedia considerando todos los árboles en el bosque. Por otro lado, para determinar la relevancia mediane RL, se utilizó la magnitud del coeficiente medio calculada a través de las réplicas de validación cruzada dividida entre la desviación estándar. Finalmente, mediante la corrida de los experimentos se

identificaron atributos en común entre todos los conjuntos de datos mediante los diferentes métodos implementado, entre ellos atributos tales como nugent, prevotella, CG1, pH, CG4, lactobacillus 2, Suterella, Ureaplasma, Coriobacteriaceae, Streptococcus, micoplasma, entre otros.

En una ampliación de investigación, los autores Beck y Foster [6] aplicaron algoritmos como *random forests* (*RF*) y *regresión logística* (*RL*) para evaluar la habilidad de los clasificadores para identificar instancias entre diferentes clases. Como parte de sus procedimientos particulares, estos métodos crean rankings de atributos basados en métricas como *purity increase in the node* por parte de *RF* y *mean coefficient magnitude* por parte de *RL.* Como parte de los experimentos, el algoritmo *relief* también fue implementado para calcular un tercer ranking de atributos. Finalmente, como parte de los resultados, se provee una tabla con los atributos más relevantes obtenidos. En dicha tabla se aprecian atributos tales como *Aeroccocus*, *Atopobium*, *Dialister*, *Eggerthella*, and *Gardnerella* entre los de mayor relevancia.

Capítulo 4. Rankings de atributos de vaginosis bacteriana mediante métodos de selección de atributos

En este capítulo se presenta el conjunto de datos utilizado en los experimentos y las fases experimentales para el cálculo de los rankings individuales y ranking generales de atributos de este proyecto de investigación. También se muestran los resultados de cada método selector de atributos implementado. Estos resultados se muestran de acuerdo al mismo orden en que fueron presentados los métodos en el capítulo 2.4.

4.1 Conjunto de datos de vaginosis bacteriana

El conjunto de datos utilizado para los experimentos de esta investigación se basa en un estudio de diagnóstico molecular de vaginosis bacteriana [88] [72]. Consiste en información clínica y microbiológica de 201 instancias y 57 atributos. Los organismos

implicados en las muestras vaginales fueron determinados mediante la técnica *Quantitative Polymerase Chain Reaction (qPCR).* Las muestras y el análisis microbiológico se realizaron en los Laboratorios de Investigación en Enfermedades Infecciosas y Metabólicas de la División Académica Multidisciplinaria de Comalcalco, Tabasco. Un resumen de los atributos que conforman la base de datos se muestra en la Tabla 2.

Tabla 2. Atributos del conjunto de datos de vaginosis bacteriana (VB) [88] .

Atributo	**Valores**
VBPCR	Etiqueta clase: 1=positivo, 2=Negativo, 3=Indeterminado
EDADENA, EDAD30	Edad del paciente
Citolog, CitologiaOrd, CitologiaBICAT	Citologia normal, ordinaria o anormal
Crispatus, L. Gasseri, L. Iners, L. Jensenii, CripatusCq, GasseriCq, JenseniiCq, InersCq, Megasphaera Phylotipo1, Atopobium, Gardnerella V., CT, NG, MH, UP, UU	Microorganismos obtenidos mediante qPCR.
BVNumero	Número de patógenos
BVCombination	Combinación de patógenos
HSV12	Herpes tipo 1
RMY0911ELSY	Relacionado con MYDE0911-ELSY
ELSY, HPV, HPVgenotypes, SingleHPVComplete, MultipleHPVComplete, LRIHPVComplete, PHRHPVComplete, HRHPVComplete	Relacionados con VPH
@6, @11, @42, @44, @84, @E626	Relacionados con VPH
E653, E666, E616, E618, E631, E633, E635, E639, E645, E651, E652, E656, E658, E659, E673	Relacionados con VPH

4.1.1 Preprocesamiento del conjunto de datos

El conjunto de datos original contiene datos nulos o faltantes (véase la Figura 4), es decir, datos que no contienen información válida. Además, contiene información relacionada al VPH que resulta ser irrelevante para el propósito de esta investigación acerca de la vaginosis bacteriana.

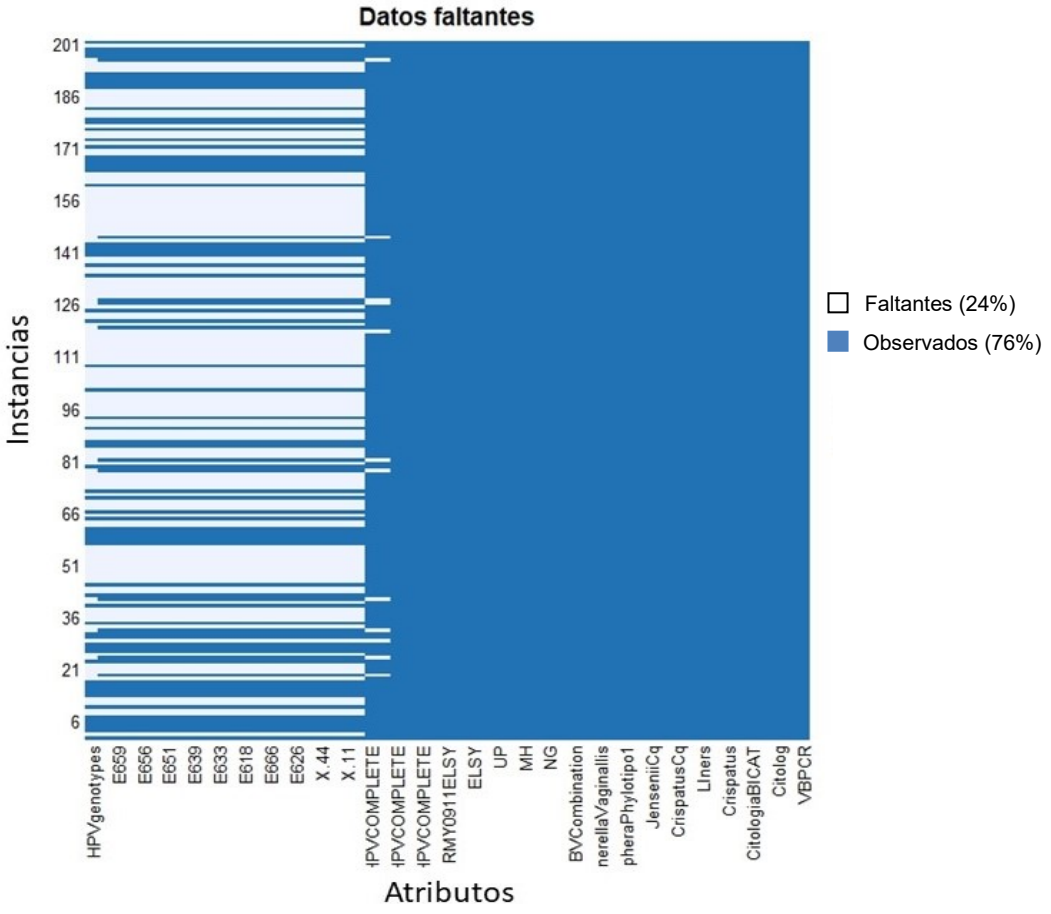

Figura 4.Representación gráfica de los datos faltantes en el conjunto de datos de la vaginosis bacteriana.

Por esta razón, antes de ser utilizado, el conjunto de datos fue preprocesado. Para ello, se llevaron a cabo las siguientes modificaciones al conjunto de datos:

1. Los atributos con información no relacionada a la VB fueron eliminados.

2. Las instancias con valores nulos, que corresponden al 24% de los datos originales, fueron eliminados.
3. La tercera clase llamada “indeterminada” fue eliminada, puesto que representa información que mediante los ensayos clínicos no pudieron ser catalogados como positivos o negativos. Esto, con la finalidad que los métodos, tanto de clasificación como de selección de atributos, identifiquen solamente entre dos clases.
4. La representación de la clase negativa 2 se cambió a 0. En el área de ciencias de datos es usual representar las etiquetas de clase de esta manera. Así, la vaginosis bacteriana positiva y negativa se identifican como 1 y 0 respectivamente.

Finalmente, el conjunto de datos preprocesado de vaginosis bacteriana contiene 173 instancias, 34 atributos y 0% de datos nulos.

4.2 Rankings individuales de atributos

La fase experimental con los algoritmos de selección de atributos consistió de 30 corridas de cada método bajo un esquema de validación cruzada de 10 pliegues. Los métodos selectores se aplicaron al conjunto de datos de entrenamiento de cada validación cruzada - representada por el 90% de los datos - el cual varia en cada iteración -pliegues-. Las 10 iteraciones de cada validación cruzada connotan los resultados de una corrida del método. Mediante las 30 corridas se calcularon 30 valores de relevancia (VR) para cada atributo de acuerdo a las métricas de cada método. Posteriormente, los 30 VR se promediaron, y se determinó el Valor de Relevancia Medio (VRM) para cada atributo. Este procedimiento se ilustra en la Figura 5.

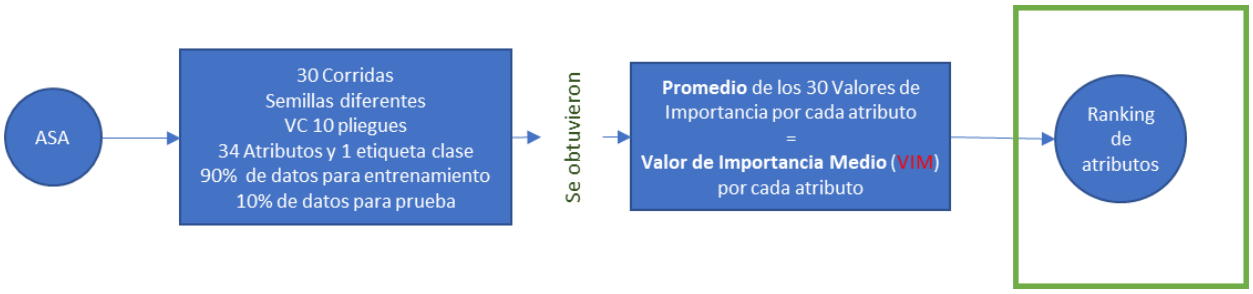

Figura 5. Diseño experimental para obtener los rankings individuales de atributos mediante las corridas de los métodos de selección de atributos. ASA: algoritmo de selección de atributos; VC: validación cruzada; VRM: valor de relevancia medio.

En Algoritmo 4 se muestra el pseudocódigo general implementado para el cálculo de los rankings individuales de atributos.

Algoritmo 4. Seudocódigo general para la corrida de los experimentos con los algoritmos de selección de atributos.

```
1. conjuntodeDatos = "vaginosisBacteriana"
2. (30 corridas)  {
3.       InicializacionSemillas ( )
4.       numerodePliegues = 10
5.            ValidacionCruzada ( conjuntodeDatos, numerodePliegue ) {
6.                 subconjuntodeEntrenamiento = 90% del conjuntodeDatos
7.                 subconjuntodePrueba = 10% del conjuntodeDatos
8.                 algoritmoSelectordeAtributos ( subconjuntodeEntrenamiento )
10.           }
11.           promedioValidacionCruzada=VC
12. }
13. promedio30corridas=VRM
```

En la Tabla 3 se muestra la estructura utilizada para promediar los 30 valores de relevancia obtenidos, al que se le denominó Valor de Relevancia Media (VRM). Esta medida, es la base para crear los rankings individuales de atributos.

Tabla 3. Estructura para la obtención del valor de relevancia medio (VRM) calculado para cada atributo en el conjunto de datos. VRM representa el promedio a través de las 30 corridas. Este proceso se utiliza para la obtención de los rankings individuales de atributos.

Atributos	**Corrida 1**	**Corrida 2**	**Corrida 3**	**…**	**Corrida *N***	**VRM**
Atributo 1	Relevancia	Relevancia	Relevancia	…	Relevancia	VRM del atributo 1
Atributo 2	Relevancia	Relevancia	Relevancia	…	Relevancia	VRM del atributo 2
Atributo 3	Relevancia	Relevancia	Relevancia	…	Relevancia	VRM del atributo 3
Atributo 4	Relevancia	Relevancia	Relevancia	…	Relevancia	VRM del atributo 4
…	…	…	…	…	…	…
Atributo *N*	Relevancia	Relevancia	Relevancia	…	Relevancia	VRM del atributo *N*

En el caso de los métodos selectores que proveen como salida subconjuntos de atributos (comúnmente los métodos tipo envoltura), el ranking individual de atributos se obtuvo mediante a un análisis de frecuencias. Éste, se basa en el cálculo de la moda estadística de las posiciones conseguidas por los atributos a través de las 30 ejecuciones de los algoritmos. En la Tabla 4 se muestra la estructura utilizada para este procedimiento.

Tabla 4. Estructura para el cálculo de los rankings individuales de métodos selectores de atributos que crean subconjuntos de atributos.

Atributos	**Corrida 1**	**Corrida 2**	**Corrida 3**	**…**	**Corrida *N***	**Moda**
Atributo 1	Posición corrida 1	Posición corrida 2	Posición corrida 3	…	Posición corrida *N*	Moda del Atributo 1
Atributo 2	Posición corrida 1	Posición corrida 2	Posición corrida 3	…	Posición corrida *N*	Moda del Atributo 2
Atributo 3	Posición corrida 1	Posición corrida 2	Posición corrida 3	…	Posición corrida *N*	Moda del Atributo 3
Atributo 4	Posición corrida 1	Posición corrida 2	Posición corrida 3	…	Posición corrida *N*	Moda del Atributo 4
…	…	…	…	…	…	…
Atributo *N*	Posición corrida 1	Posición corrida 2	Posición corrida 3	…	Posición corrida *N*	Moda del Atributo *N*

4.2.1 *Relief*

Los experimentos realizados con el algoritmo *Relief* fueron completados. Con base en las 30 corridas del selector de atributos *Relief*, se obtuvo un valor de relevancia medio (VRM) al promediar los 30 valores de relevancia obtenidos por los atributos de la

vaginosis bacteriana en cada corrida con validación cruzada. La manera en que el método evalúa la relevancia de cada atributo en el conjunto de datos de VB se detalla en la Sección 2.4. En la Figura 6 se muestra el VRM obtenido por cada uno de los atributos mediante *Relief*.

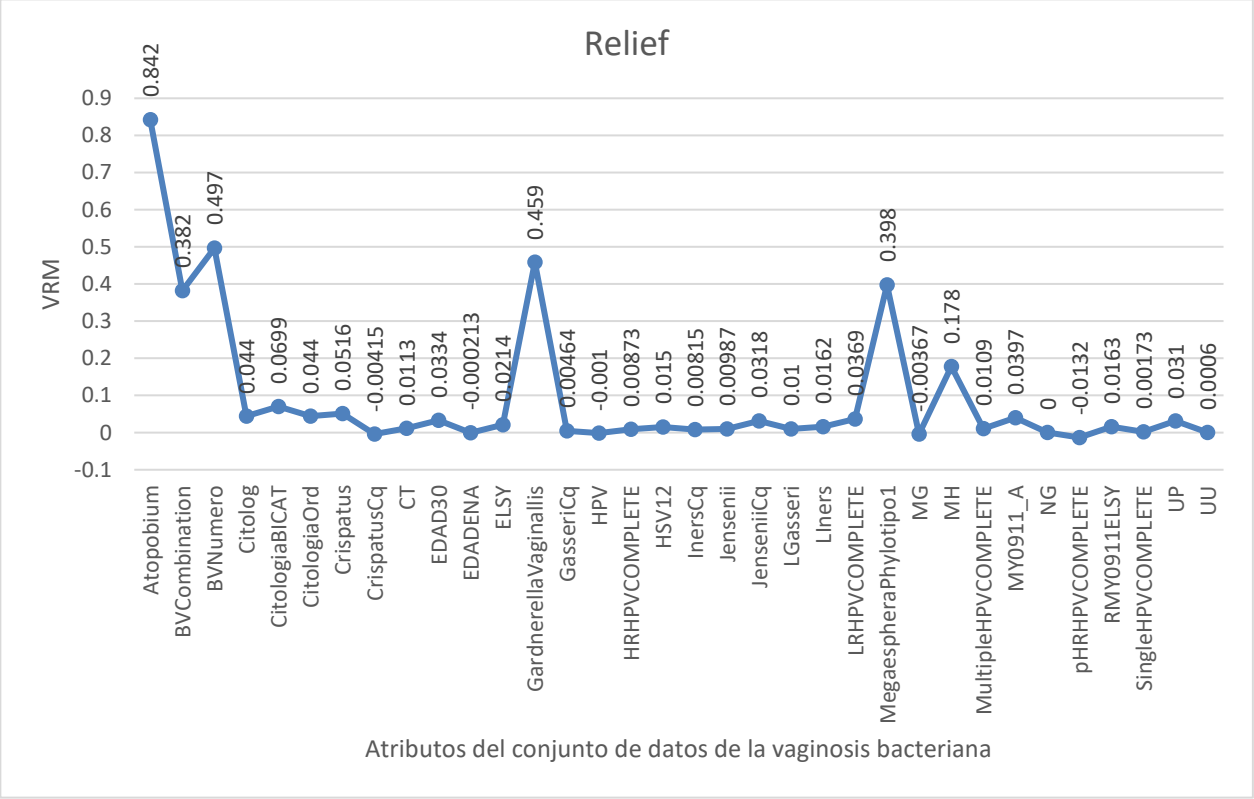

Figura 6. Valores de relevancia media (VRM) obtenida por los atributos del conjunto de vaginosis bacteriana mediante las 30 ejecuciones con validación cruzada del método de selección de atributos *Relief*

A partir de este procedimiento, y al ordenar los atributos en forma descendente de acuerdo al VRM obtenido mediante *Relief*, se crea el denominado “ranking individual de atributos”. Dicho ranking se muestra en la Tabla 5.

Tabla 5. Ranking individual de atributos de vaginosis bacteriana generado por el algoritmo *relief* mediante las 30 corridas bajo esquema de validación cruzada. Los atributos se ordenan de forma descendente de acuerdo al Valor de Relevancia Medio (VRM) -promedio de los 30 resultados-.

#Ranking	Features	VRM	#Ranking	Atributos	VRM
1	Atopobium	0.842	18	LIners	0.0162
2	BVNumero	0.497	19	HSV12	0.015
3	GardnerellaVaginallis	0.459	20	CT	0.0113
4	MegaespheraPhylotipo1	0.398	21	MultipleHPVCOMPLETE	0.0109
5	BVCombination	0.382	22	LGasseri	0.01
6	MH	0.178	23	Jensenii	0.00987
7	CitologiaBICAT	0.0699	24	HRHPVCOMPLETE	0.00873
8	Crispatus	0.0516	25	InersCq	0.00815
9	Citolog	0.044	26	GasseriCq	0.00464
10	CitologiaOrd	0.044	27	SingleHPVCOMPLETE	0.00173
11	MY0911_A	0.0397	28	UU	0.0006
12	LRHPVCOMPLETE	0.0369	29	NG	0
13	EDAD30	0.0334	30	EDADENA	-0.000213
14	JenseniiCq	0.0318	31	HPV	-0.001
15	UP	0.031	32	MG	-0.00367
16	ELSY	0.0214	33	CrispatusCq	-0.00415
17	RMY0911ELSY	0.0163	34	pHRHPVCOMPLETE	-0.0132

En este caso, *Relief* identificó a “Atopobium” como el atributo más relevante de la VB. Además, atributos como “BVNumero”, “GardnerellaVaginallis”, “MegaespheraPhylotipo1”, “BVCombination” y “MH” obtuvieron valores de relevancia por encima de la media aritmética (0.95).

4.2.2 Chi cuadrado

La identificación de los atributos más relevantes para el diagnóstico de la vaginosis bacteriana mediante el método chi cuadrado fue implementado de la siguiente manera. Se realizaron 30 corridas del método bajo un esquema de validación cruzada de 10 pliegues. En cada iteración del esquema de validación cruzada se utilizó el 90% de las instancias para crear un conjunto de entrenamiento, la cual es aleatoriamente cambiado en cada iteración. Por cada una de las iteraciones, se calculó en nivel de relevancia de cada atributo del conjunto de entrenamiento basado en el chi cuadrado. En cada corrida, se obtuvo un valor de relevancia para cada atributo al promediar las 10 medidas obtenidas en las iteraciones. Cada corrida fue implementada con semillas diferentes

para la aleatoriedad de los datos. Al final, las 30 medidas fueron promediadas, con lo cual, se obtuvo un valor de relevancia medio (VRM). Este valor es utilizado como criterio de ranqueo. En la Figura 7 se muestran los VRM resultantes de cada atributo obtenido mediante la medida estadística chi cuadrado.

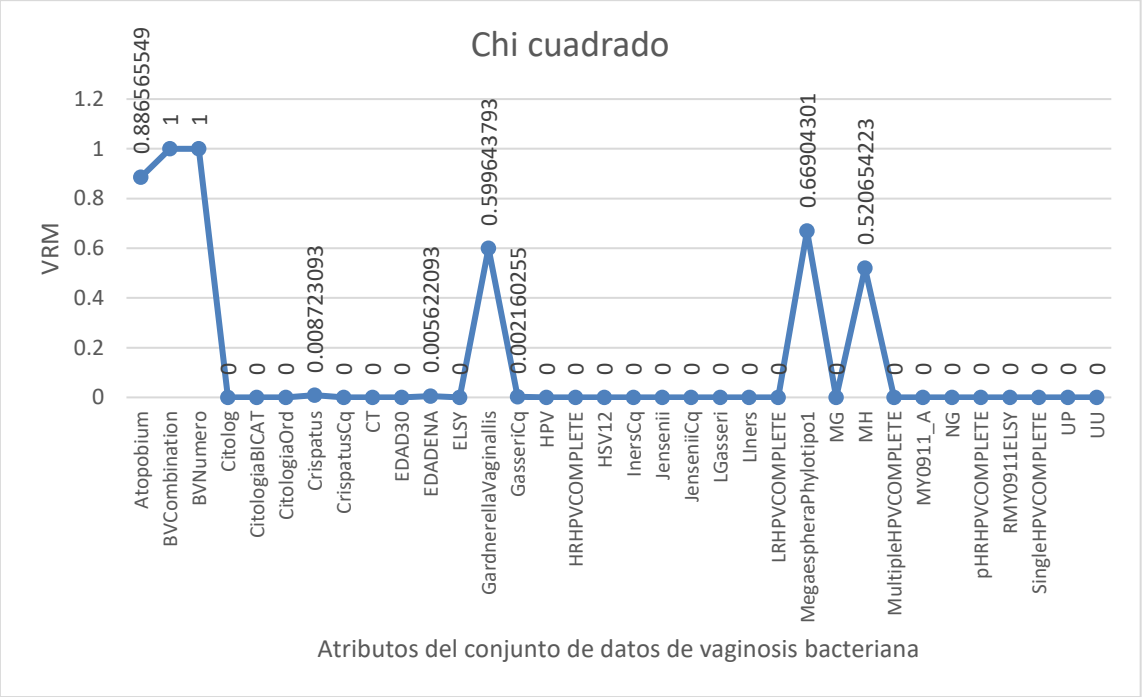

Figura 7. Valores de relevancia media (VRM) obtenida por los atributos del conjunto de vaginosis bacteriana mediante las 30 ejecuciones con validación cruzada del método de selección de atributos *chi cuadrado.*

Con base en estos resultados, se crea el ranking individual de atributos. Para ello, los atributos en el conjunto de datos son ordenados de forma descendente respecto al VRM obtenido de las 30 corridas. El ranking individual de atributos de la vaginosis bacteriana basado en la medida de chi cuadrada se muestra en la Tabla 6.

Tabla 6. Ranking individual de atributos de vaginosis bacteriana generado mediante la medida estadística *chi cuadrado*. Los atributos se ordenan de forma descendente de acuerdo al Valor de Relevancia Medio (VRM) -promedio de los 30 resultados-.

#Rank	Features	VRM	#Rank	Features	VRM
1	BVCombination	1	18	HRHPVCOMPLETE	0
2	BVNumero	1	19	HSV12	0
3	Atopobium	0.88656555	20	InersCq	0
4	MegaespheraPhylotipo1	0.66904301	21	Jensenii	0
5	GardnerellaVaginallis	0.59964379	22	JenseniiCq	0
6	MH	0.52065422	23	LGasseri	0
7	Crispatus	0.00872309	24	LIners	0
8	EDADENA	0.00562209	25	LRHPVCOMPLETE	0
9	GasseriCq	0.00216026	26	MG	0
10	Citolog	0	27	MultipleHPVCOMPLETE	0
11	CitologiaBICAT	0	28	MY0911_A	0
12	CitologiaOrd	0	29	NG	0
13	CrispatusCq	0	30	pHRHPVCOMPLETE	0
14	CT	0	31	RMY0911ELSY	0
15	EDAD30	0	32	SingleHPVCOMPLETE	0
16	ELSY	0	33	UP	0
17	HPV	0	34	UU	0

4.2.3 Entropía

Los experimentos realizados con el algoritmo *DT* fueron completados. Con base en las 30 corridas de *DT* como selector de atributos, se obtuvo un valor de relevancia medio (VRM) al promediar los 30 valores de relevancia obtenidos por los atributos de la vaginosis bacteriana en cada corrida con validación cruzada. Los detalles de cómo el algoritmo evalúa la relevancia de cada atributo en el conjunto de datos de VB con base en la entropía se detalla en la Sección 2.4.

Los resultados de los experimentos con *DT* se muestran a continuación. Con base en los resultados obtenidos por la implementación del modelo de árboles de decisión, puede ser obtenido un gráfico en forma de árbol que representa la o las reglas creadas por el método clasificador. Este grafico es mostrado en la Figura 8.

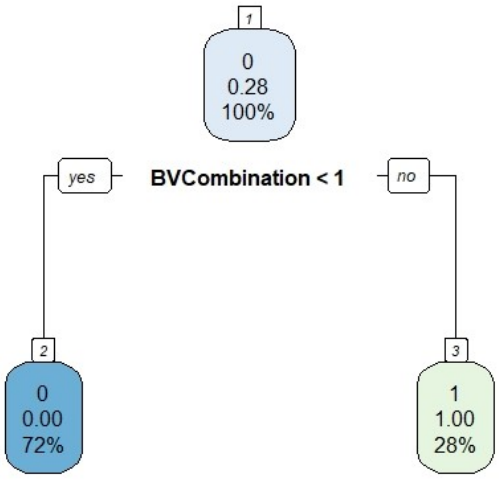

Figura 8. Representación gráfica de un *árbol de decisión* a partir de las reglas creadas por el método del mismo nombre.

En Figura 8 se muestra cómo el modelo basado en arboles de decisión definió como el nodo raíz al atributo “BVCombination”. La evaluación del modelo clasificador con este atributo fue más que necesario para crear una sola regla de clasificación:

if BVCombination < 1 then VBPCR = 0 (VB negativo)
If BVCombination >= 1 then VBPCR = 1 (VB positivo)

A pesar que el método *DT* con base en la medida de *entropía* definió un solo atributo como el de mayor relevancia, en esta investigación es de interés la creación de rankings de atributos. Por tanto, con base en el modelo creado con *DT* fue extraído el valor de relevancia medio de cada atributo a través de las corridas de entrenamiento. La Figura 9 muestra los VRM obtenido por los atributos mediante el cálculo de la entropía.

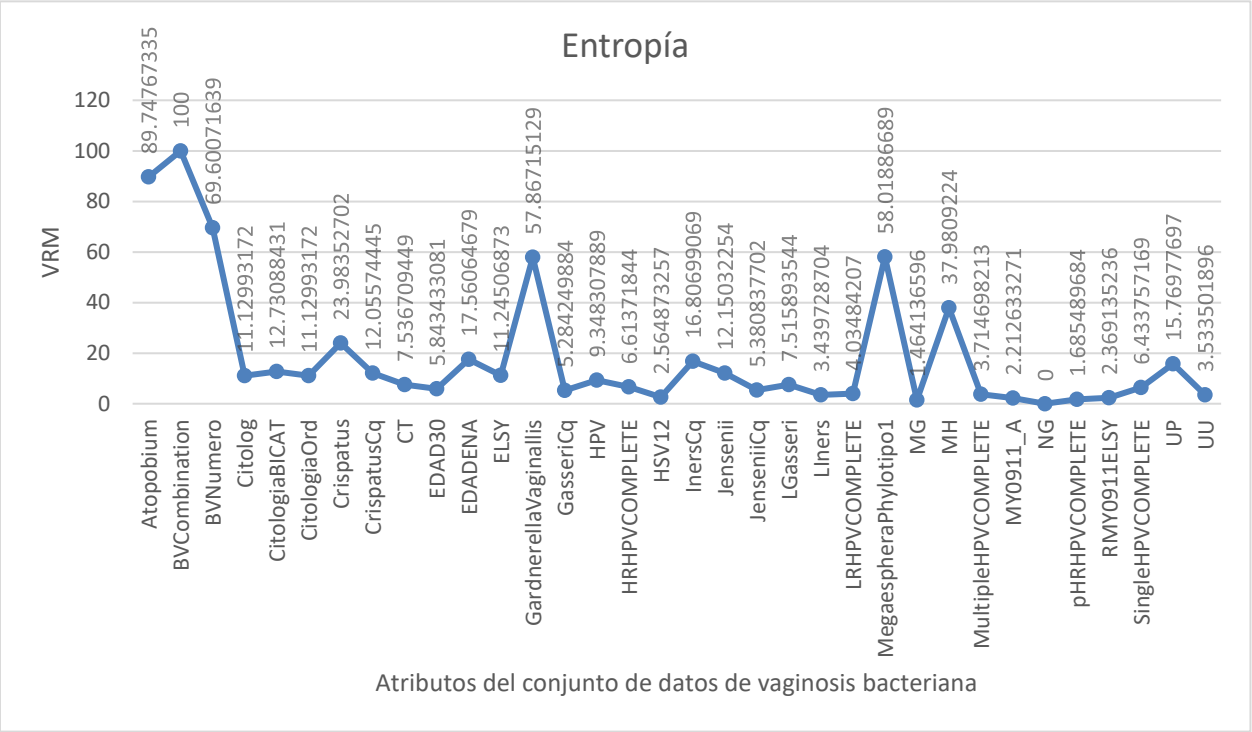

Figura 9. Valores de relevancia media (VRM) obtenida por los atributos del conjunto de vaginosis bacteriana mediante las 30 ejecuciones con validación cruzada del método de selección de atributos *Relief*

Con base en estos resultados, se generó el ranking individual de atributos utilizando *DT* como método selector de atributos. Dicho ranking se muestra en la Tabla 7.

Tabla 7. Ranking individual de atributos de vaginosis bacteriana generado por el método *DT* con base en la *Entropía* mediante las 30 corridas bajo esquema de validación cruzada. Los atributos se ordenan de forma descendente de acuerdo al Valor de Relevancia Medio (VRM) -promedio de los 30 resultados-.

#Ranking	Features	VRM	#Ranking	Atributos	VRM
1	BVCombination	100	18	CT	7.53670945
2	Atopobium	89.7476733	19	LGasseri	7.51589354
3	BVNumero	69.6007164	20	HRHPVCOMPLETE	6.61371844
4	MegaespheraPhylotipo1	58.0188669	21	SingleHPVCOMPLETE	6.43375717
5	GardnerellaVaginallis	57.8671513	22	EDAD30	5.84343308
6	MH	37.9809224	23	JenseniiCq	5.3808377
7	Crispatus	23.983527	24	GasseriCq	5.28424988
8	EDADENA	17.5606468	25	LRHPVCOMPLETE	4.03484207
9	InersCq	16.8069907	26	MultipleHPVCOMPLETE	3.71469821
10	UP	15.769777	27	UU	3.5335019
11	CitologiaBICAT	12.7308843	28	LIners	3.4397287
12	Jensenii	12.1503225	29	HSV12	2.56487326
13	CrispatusCq	12.0557445	30	RMY0911ELSY	2.36913524
14	ELSY	11.2450687	31	MY0911_A	2.21263327
15	CitologiaOrd	11.1299317	32	pHRHPVCOMPLETE	1.68548968
16	Citolog	11.1299317	33	MG	1.4641366
17	HPV	9.34830789	34	NG	0

4.2.4 Ganancia de información

El ranking con los atributos más relevantes para el diagnóstico de la vaginosis bacteriana mediante el método basado en la ganancia de información fue calculado de la siguiente manera. Se realizaron 30 corridas del método bajo un esquema de validación cruzada de 10 pliegues. En cada iteración del esquema de validación cruzada se utilizó el 90% de las instancias para crear un conjunto de entrenamiento, el cual fue utilizado para la implementación del método. Las instancias de entrenamiento son aleatoriamente cambiadas en cada iteración. Por cada una de las iteraciones, se calculó en nivel de relevancia de cada atributo del conjunto de entrenamiento basado en el criterio de ranqueo basado en la ganancia de información descrito anteriormente en la sección 2.4. En cada corrida, se obtuvo un valor de relevancia para cada atributo al promediar las 10 medidas obtenidas en las iteraciones. Al final, las 30 medidas de relevancia fueron promediadas, con lo cual se obtuvo un valor de relevancia medio

(VRM). Este valor es utilizado como criterio de ranqueo. Cada corrida fue implementada con semillas diferentes para asegurar la aleatoriedad de los datos En la Figura 10 se muestran los VRM resultantes de cada atributo obtenido mediante la medida “Ganancia de información”.

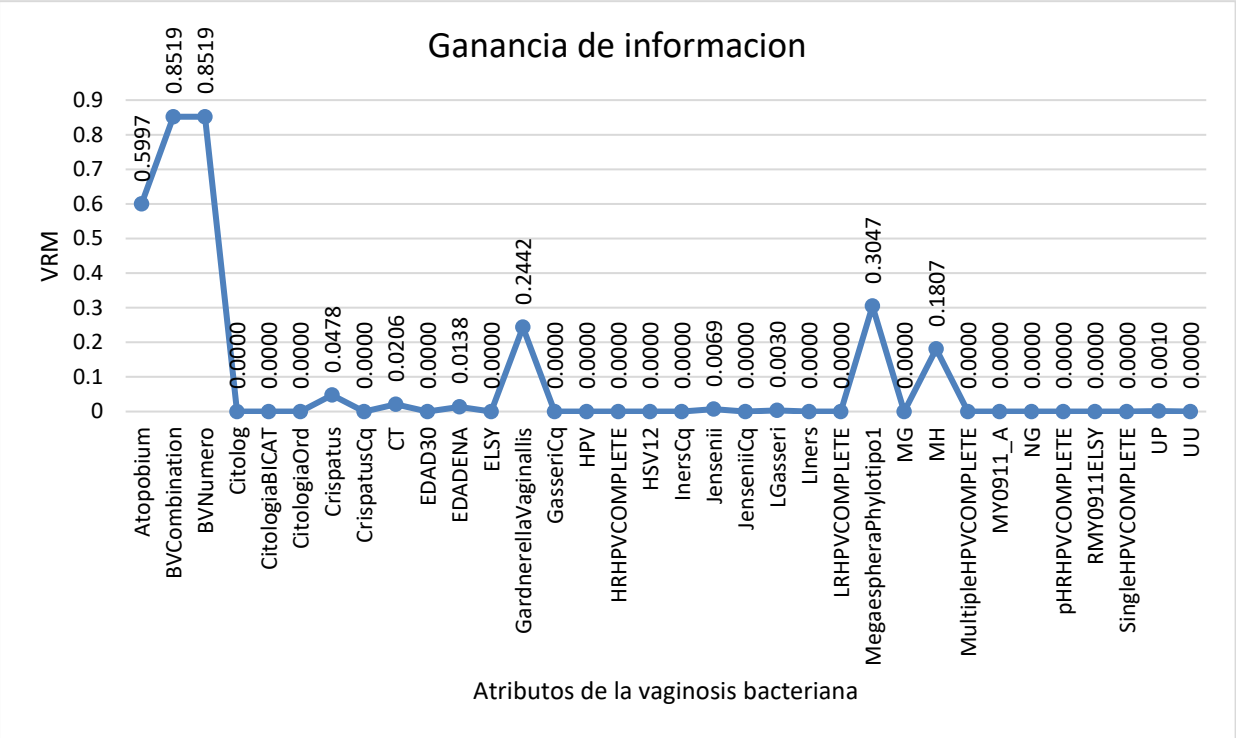

Figura 10.. Valores de relevancia media (VRM) obtenida por los atributos del conjunto de vaginosis bacteriana mediante las 30 ejecuciones con validación cruzada del método de selección de atributos *DT* basado en la medida de ganancia de información.

Con base en estos resultados, se crea el ranking individual de atributos. Para ello, los atributos en el conjunto de datos son ordenados de forma descendente respecto al VRM obtenido de las 30 corridas. El ranking individual de atributos de la vaginosis bacteriana basado en el método y medida ganancia de información es mostrado en la Tabla 8.

Tabla 8. Ranking individual de atributos de vaginosis bacteriana generado mediante el algoritmo *Information.gain*. Los atributos se ordenan de forma descendente de acuerdo al Valor de Relevancia Medio (VRM) -promedio de los 30 resultados-.

#Rank	Features	VRM	#Rank	Features	VRM
1	BVCombination	0.85194553	18	ELSY	0
2	BVNumero	0.85194553	19	GasseriCq	0
3	Atopobium	0.599699	20	HPV	0
4	MegaespheraPhylotipo1	0.30470137	21	HRHPVCOMPLETE	0
5	GardnerellaVaginallis	0.2441609	22	HSV12	0
6	MH	0.18067268	23	InersCq	0
7	Crispatus	0.04778396	24	JenseniiCq	0
8	CT	0.02061765	25	LIners	0
9	EDADENA	0.01379158	26	LRHPVCOMPLETE	0
10	Jensenii	0.00688699	27	MG	0
11	LGasseri	0.00303718	28	MultipleHPVCOMPLETE	0
12	UP	0.00098758	29	MY0911_A	0
13	Citolog	0	30	NG	0
14	CitologiaBICAT	0	31	pHRHPVCOMPLETE	0
15	CitologiaOrd	0	32	RMY0911ELSY	0
16	CrispatusCq	0	33	SingleHPVCOMPLETE	0
17	EDAD30	0	34	UU	0

En el caso de este método basado en la medida denominada ganancia de informacion, se identificaron 12 de 34 atributos como relevantes para la vaginosis bacteriana. Sin embargo, sólo 6 están por encima del VRM medio (0.09194794), éstos son los atributos llamados *BVCombination, BVNumero, Atopobium, MegaespheraPhylotipo1, GardnerellaVaginallis y MH.*

4.2.5 Incertidumbre simétrica

Los experimentos para la identificación de los atributos más relevantes en el diagnóstico de la vaginosis bacteriana mediante el método de *incertidumbre simétrica* fueron implementados de la siguiente manera. Se realizaron 30 corridas del método bajo un esquema de validación cruzada de 10 pliegues. En cada iteración del esquema de validación cruzada se utilizó el 90% de las instancias para crear un conjunto de entrenamiento, la cual es aleatoriamente cambiado en cada iteración. Por cada una de las iteraciones, se calculó en nivel de relevancia de cada atributo del conjunto de entrenamiento con base en el método *incertidumbre simétrica* tal como se describió anteriormente en la sección 2.1. En cada corrida, se obtuvo un valor de relevancia para cada atributo al promediar las 10 medidas obtenidas en las iteraciones. Al final, las 30 medidas de relevancia se promediaron, con lo cual, se obtuvo un valor de relevancia medio (VRM). Cada corrida fue implementada con semillas diferentes para asegurar la aleatoriedad de los datos. En la Figura 11 se muestran los VRM resultantes de cada atributo obtenido a través del cálculo de los pesos de los atributos mediante *incertidumbre simétrica.*

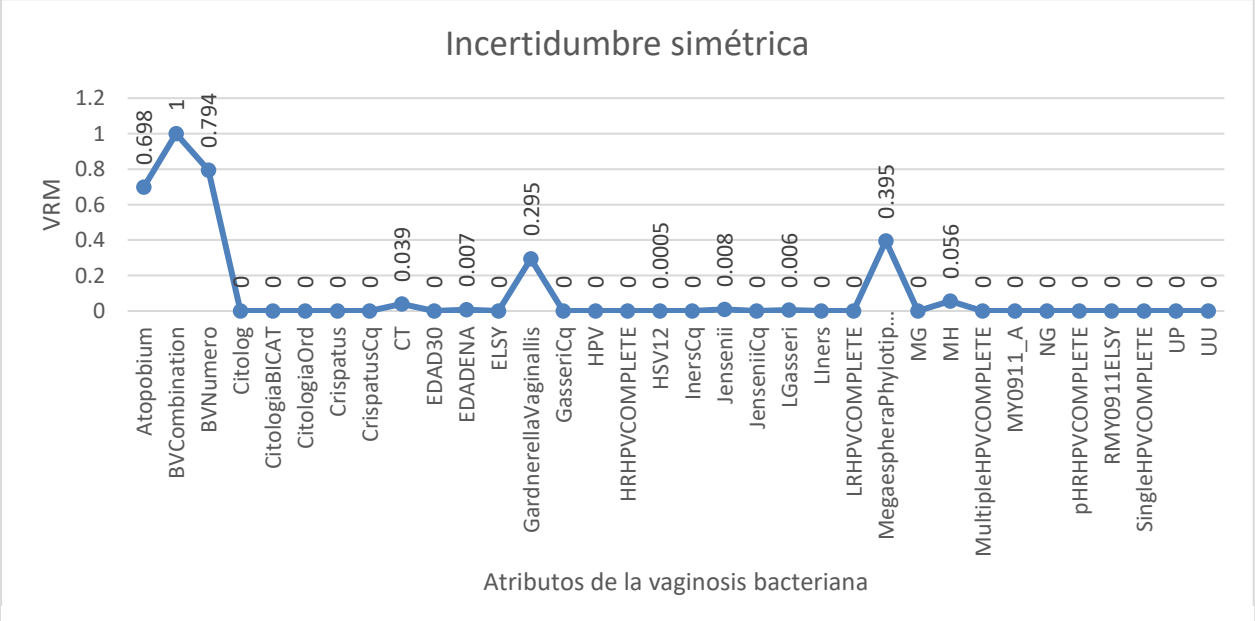

Figura 11. Valor de relevancia media (VRM) obtenida por los atributos de la vaginosis bacteriana en las 30 corridas del método *incertidumbre simétrica.*

Con base en los resultados obtenidos anteriormente, se crea el ranking individual de atributos mediante el método *incertidumbre simétrica.* Para ello, los atributos en el conjunto de datos son ordenados de forma descendente al considerar el VRM obtenido de las 30 corridas como criterio de ordenamiento. El ranking individual de atributos de la vaginosis bacteriana basado en el método *incertidumbre simétrica* se muestra en la Tabla 9.

Tabla 9. Ranking individual de atributos de vaginosis bacteriana generado mediante el método *incertidumbre simétrica*. Los atributos se ordenan de forma descendente de acuerdo al Valor de Relevancia Medio (VRM) -promedio de los 30 resultados-.

#Rank	Atributos	VRM	#Rank	Atributos	VRM
1	BVCombination	1	18	ELSY	0
2	BVNumero	0.794	19	GasseriCq	0
3	Atopobium	0.698	20	HPV	0
4	MegaespheraPhylotipo1	0.395	21	HRHPVCOMPLETE	0
5	GardnerellaVaginallis	0.295	22	InersCq	0
6	MH	0.056	23	JenseniiCq	0
7	CT	0.039	24	LIners	0
8	Jensenii	0.008	25	LRHPVCOMPLETE	0
9	EDADENA	0.007	26	MG	0
10	LGasseri	0.006	27	MultipleHPVCOMPLETE	0
11	HSV12	0.0005	28	MY0911_A	0
12	Citolog	0	29	NG	0
13	CitologiaBICAT	0	30	pHRHPVCOMPLETE	0
14	CitologiaOrd	0	31	RMY0911ELSY	0
15	Crispatus	0	32	SingleHPVCOMPLETE	0
16	CrispatusCq	0	33	UP	0
17	EDAD30	0	34	UU	0

En este caso, el método basado en *incertidumbre simétrica* identificó a “BVCombination” como el atributo más relevante de la VB. Además, atributos como “BVNumero”, “Atopobium”, “MegaespheraPhylotipo1” y “GardnerellaVaginallis” obtuvieron valores de relevancia por encima de la media aritmética (0.097).

4.2.6 Longitud de descripción media (MDL)

La identificación de los atributos más relevantes para el diagnóstico de la vaginosis bacteriana mediante el algoritmo *MDL* se obtuvieron de la siguiente manera. Se realizaron 30 corridas del método bajo un esquema de validación cruzada de 10 pliegues. En cada iteración del esquema de validación cruzada se utilizó el 90% de las instancias para crear un conjunto de entrenamiento, el cual es aleatoriamente cambiado en cada iteración. Por cada una de las iteraciones, se calculó el nivel de relevancia de cada atributo del conjunto de entrenamiento con base en el método *MDL* tal como se describió anteriormente. En cada corrida, se obtuvo un valor de relevancia para cada atributo al promediar las 10 medidas obtenidas en las iteraciones. Al final, las 30 medidas de relevancia se promediaron, con lo cual se obtuvo un valor de relevancia medio (VRM). Cada corrida fue implementada con semillas diferentes para asegurar la aleatoriedad de los datos. En la Figura 12 se muestran los VRM resultantes de cada atributo obtenido a través de *MDL*.

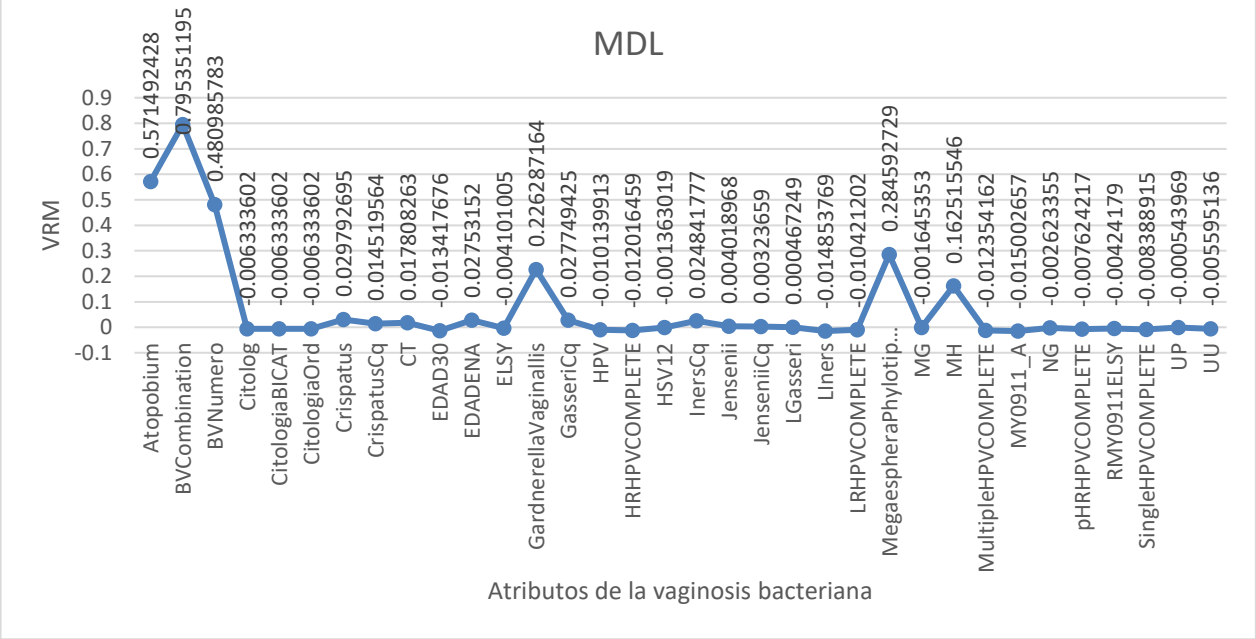

Figura 12. Valor de relevancia media (VRM) obtenida por los atributos de la vaginosis bacteriana en las 30 corridas del método *MDL*.

Con base en los resultados obtenidos, se crea el ranking individual de atributos mediante el método *MDL.* Para ello, los atributos en el conjunto de datos son ordenados de forma descendente al considerar el VRM obtenido de las 30 corridas como criterio de ordenamiento. El ranking individual de atributos de la vaginosis bacteriana basado en el método *MDL* se muestra en la Tabla 10.

Tabla 10. Ranking individual de atributos de vaginosis bacteriana generado mediante el método MDL. Los atributos se ordenan de forma descendente de acuerdo al Valor de Relevancia Medio (VRM) -promedio de las 30 corridas del método bajo VC-10-.

#Rank	Atributo	VRM	#Rank	Atributo	VRM
1	BVCombination	0.7953	18	MG	0
2	Atopobium	0.5714	19	NG	-0.0026
3	BVNumero	0.4809	20	ELSY	-0.0041
4	MegaespheraPhylotipo1	0.2845	21	RMY0911ELSY	-0.0042
5	GardnerellaVaginallis	0.2262	22	UU	-0.0055
6	MH	0.1625	23	Citolog	-0.0063
7	Crispatus	0.0297	24	CitologiaBICAT	-0.0063
8	GasseriCq	0.0277	25	CitologiaOrd	-0.0063
9	EDADENA	0.0275	26	pHRHPVCOMPLETE	-0.0076
10	InersCq	0.0248	27	SingleHPVCOMPLETE	-0.0083
11	CT	0.0178	28	HPV	-0.0101
12	CrispatusCq	0.0145	29	LRHPVCOMPLETE	-0.0104
13	Jensenii	0.0040	30	HRHPVCOMPLETE	-0.0120
14	JenseniiCq	0.0032	31	MultipleHPVCOMPLETE	-0.0123
15	LGasseri	0.0004	32	EDAD30	-0.0134
16	UP	-0.0005	33	LIners	-0.0148
17	HSV12	-0.0013	34	MY0911_A	-0.0150

En este caso, si se considera la media aritmética (0.074) como punto de partida para definir los atributos más importantes de la VB mediante el método MDL, encontramos atributos como “BVCombination”, “Atopobium”, “BVNumero”, “MegaespheraPhylotipo1”, “GardnerellaVaginallis” y “MH”.

4.2.7 Puntaje de Fisher

La identificación de los atributos más relevantes para el diagnóstico de la vaginosis bacteriana mediante el algoritmo *puntaje de Fisher* se obtuvieron de la siguiente manera. Se realizaron 30 corridas del método bajo un esquema de validación cruzada de 10 pliegues. En cada iteración del esquema de validación cruzada se utilizó el 90% de las instancias para crear un conjunto de entrenamiento, el cual es cambiado aleatoriamente en cada iteración. Por cada una de las iteraciones, se calculó en nivel de relevancia de cada atributo del conjunto de entrenamiento con base en el método *puntaje de Fisher* tal como se describió anteriormente. Al final, las 30 medidas de relevancia se promediaron, con lo cual se obtuvo un valor de relevancia medio (VRM). Cada corrida fue implementada con semillas diferentes para asegurar la aleatoriedad de los datos. En la Figura 13 se muestran los VRM resultantes de cada atributo obtenido a través del *puntaje de Fisher*.

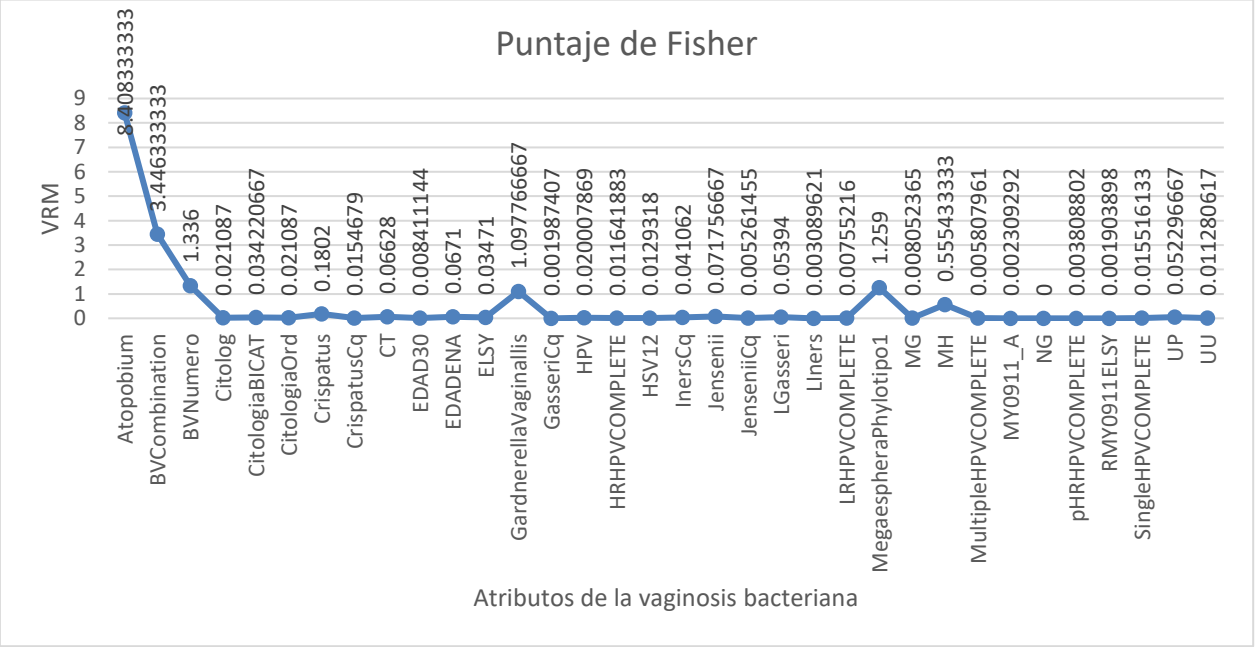

Figura 13. Valor de relevancia media (VRM) obtenida por los atributos de la vaginosis bacteriana en las 30 corridas del método *puntaje de Fisher.*

Con base en los resultados obtenidos, se crea el ranking individual de atributos mediante el método *puntaje de Fisher.* Para ello, los atributos en el conjunto de datos son ordenados de forma descendente al considerar el VRM obtenido de las 30 corridas como criterio de ordenamiento. El ranking individual de atributos de la vaginosis bacteriana basado en el método *puntaje de Fisher* se muestra en la Tabla 11.

Tabla 11. Ranking individual de atributos de vaginosis bacteriana generado mediante el método *puntaje de Fisher*. Los atributos se ordenan de forma descendente de acuerdo al Valor de Relevancia Medio (VRM) -promedio de las 30 corridas del método bajo VC-10-.

#Rank	Atributo	VRM	#Rank	Atributo	VRM
1	Atopobium	8.4083	18	HPV	0.0200
2	BVCombination	3.4463	19	SingleHPVCOMPLETE	0.0155
3	BVNumero	1.3360	20	CrispatusCq	0.0155
4	MegaespheraPhylotipo1	1.2590	21	HSV12	0.0129
5	GardnerellaVaginallis	1.0978	22	HRHPVCOMPLETE	0.0116
6	MH	0.5554	23	UU	0.0113
7	Crispatus	0.1802	24	EDAD30	0.0084
8	Jensenii	0.0718	25	MG	0.0081
9	EDADENA	0.0671	26	LRHPVCOMPLETE	0.0076
10	CT	0.0663	27	MultipleHPVCOMPLETE	0.0058
11	LGasseri	0.0539	28	JenseniiCq	0.0053
12	UP	0.0523	29	pHRHPVCOMPLETE	0.0038
13	InersCq	0.0411	30	LIners	0.0031
14	ELSY	0.0347	31	MY0911_A	0.0023
15	CitologiaBICAT	0.0342	32	GasseriCq	0.0020
16	Citolog	0.0211	33	RMY0911ELSY	0.0019
17	CitologiaOrd	0.0211	34	NG	0.0000

Mediante el método de *puntaje de Fisher*, se determinó que entre los atributos más relevantes para la VB se encuentran atributos como “Atopobium”, “BVCombination”, “BVNumero”, “MegaespheraPhylotipo1”, “GardnerellaVaginallis” y “MH”. Estos atributos obtuvieron un VRM por encima de la media aritmética (0.4965).

4.2.8 Selección de atributos basada en correlaciones.

La identificación de los atributos más relevantes para el diagnóstico de la vaginosis bacteriana mediante el algoritmo *correlated-feature selection (CFS)* se efectuó de la siguiente manera. Se realizaron 30 corridas del método bajo un esquema de validación cruzada de 10 pliegues. Cada corrida fue implementada con semillas diferentes para

asegurar la aleatoriedad de los datos. En cada iteración del esquema de validación cruzada se utilizó el 90% de las instancias para crear un conjunto de entrenamiento, la cual, es aleatoriamente cambiado en cada iteración. En cada una de las iteraciones se implementó el método *CFS* sobre el conjunto de entrenamiento. Por cada iteración se obtuvo un subconjunto de datos como resultado de la implementación del *CFS*. Este subconjunto representa los atributos identificados como más relevantes de la VB respecto al método en cuestión. Al finalizar, se obtuvieron 300 subconjuntos de datos. A partir de estos subconjuntos, se realizó un análisis de distribución de frecuencias, que permite identificar el número de veces que un atributo se encuentra entre los subconjuntos de atributos resultantes con el método *CFS*. Los resultados del análisis de frecuencia se representan en la Figura 14.

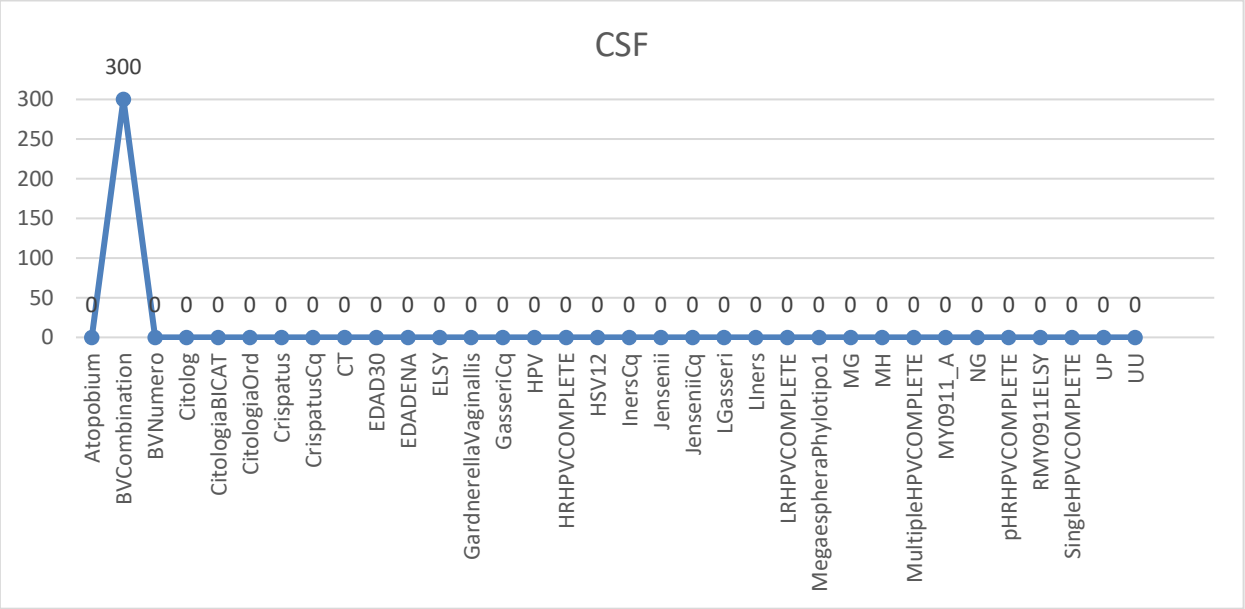

Figura 14. Distribución de frecuencias de los atributos de la vaginosis bacteriana a través de las 30 corridas con VC-10 del método selección basada en correlaciones (*CFS*, por sus siglas en inglés).

Con base en los resultados obtenidos, se crea el ranking individual de atributos mediante el método *CFS.* Para ello, los atributos en el conjunto de datos de VB son ordenados de forma descendente al considerar la frecuencia obtenida de las 30 corridas

como criterio de ordenamiento. El ranking individual de atributos de la vaginosis bacteriana basado en el método *CFS* se muestra en la Tabla 12

Tabla 12. Ranking individual de atributos de vaginosis bacteriana generado por el algoritmo *selección basada en correlaciones* mediante las 30 corridas bajo esquema de validación cruzada. Los atributos se ordenan de forma descendente de acuerdo a la frecuencia obtenida de los 300 resultados).

#Rank	Atributo	Frecuencia	#Rank	Atributo	Frecuencia
1	BVCombination	300	18	InersCq	0
2	Atopobium	0	19	Jensenii	0
3	BVNumero	0	20	JenseniiCq	0
4	Citolog	0	21	LGasseri	0
5	CitologiaBICAT	0	22	LIners	0
6	CitologiaOrd	0	23	LRHPVCOMPLETE	0
7	Crispatus	0	24	MegaespheraPhylotipo1	0
8	CrispatusCq	0	25	MG	0
9	CT	0	26	MH	0
10	EDAD30	0	27	MultipleHPVCOMPLETE	0
11	EDADENA	0	28	MY0911_A	0
12	ELSY	0	29	NG	0
13	GardnerellaVaginallis	0	30	pHRHPVCOMPLETE	0
14	GasseriCq	0	31	RMY0911ELSY	0
15	HPV	0	32	SingleHPVCOMPLETE	0
16	HRHPVCOMPLETE	0	33	UP	0
17	HSV12	0	34	UU	0

Nótese en este caso, que el único atributo identificado como relevante por el método *CFS* en el conjunto de datos de la VB es “BVCombination”. De las 30 corridas realizadas del método *CFS*, que permitió obtener 30 subconjuntos, el subconjunto óptimo se conforma de un solo atributo.

4.2.9 Correlación de Pearson

La identificación de los atributos más relevantes de la VB mediante el método de correlación de Pearson se efectuó siguiendo el proceso descrito en la Sección 2.4. Esto es, se realizaron 30 corridas del método bajo un esquema de validación cruzada de 10 pliegues. Como resultado, se obtuvieron 30 coeficientes de correlación para cada atributo del conjunto de datos, los cuales fueron promediados para obtener un valor de relevancia medio (VRM). Este valor representa el nivel de relevancia para la VB. Los VRM resultantes para cada atributo se representan en la Figura 15.

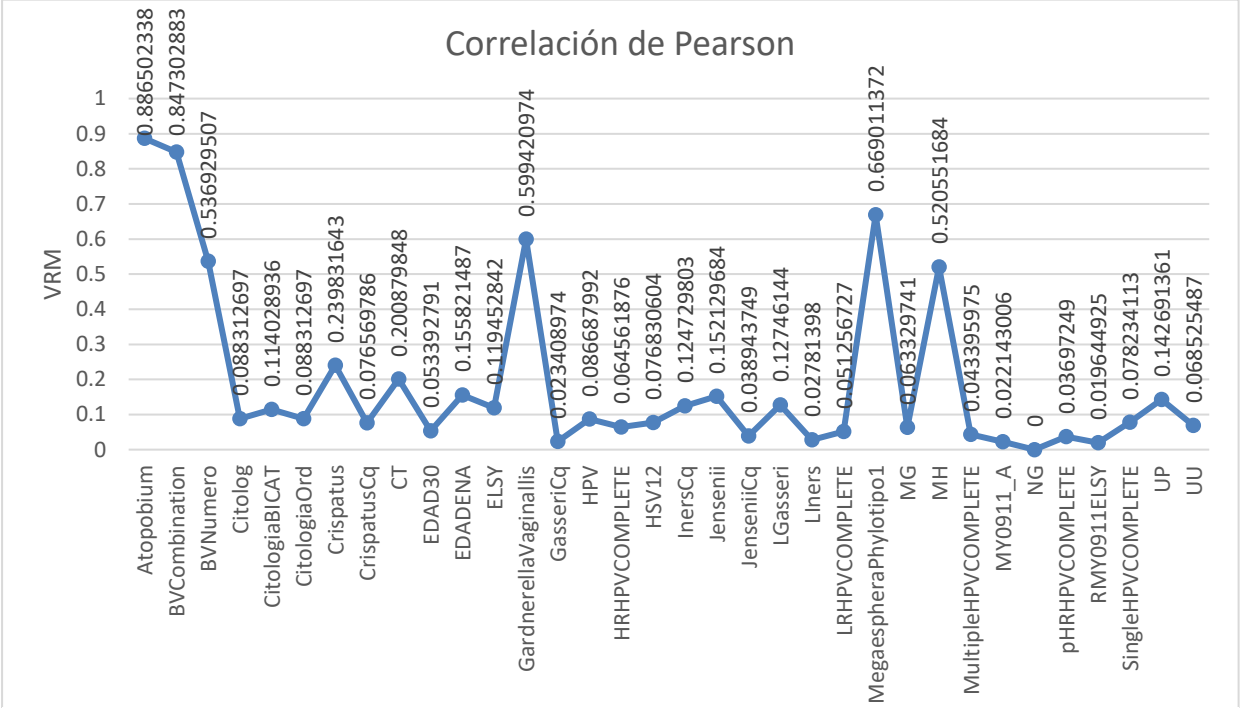

Figura 15. Valor de relevancia media (VRM) obtenida por los atributos de la vaginosis bacteriana en las 30 corridas del método *Correlación de Pearson.*

Con base en los resultados obtenidos, se crea el ranking individual de atributos bajo el criterio de *Correlación de Pearson.* Para ello, los atributos en el conjunto de datos de VB se ordenaron de forma descendente al considerar el VRM obtenida de las 30

corridas. El ranking individual de atributos de la vaginosis bacteriana basado en el método *Correlación de Pearson* se muestra en la Tabla 13.

Tabla 13. Ranking individual de atributos de vaginosis bacteriana generado por el método *Correlación de Pearson* mediante las 30 corridas bajo esquema de validación cruzada. Los atributos se ordenan de forma descendente de acuerdo al valor de relevancia media (VRM) obtenida de los resultados.

#Rank	Atributo	VRM	#Rank	Atributo	VRM
1	Atopobium	0.8865	18	HPV	0.0867
2	BVCombination	0.8473	19	SingleHPVCOMPLETE	0.0782
3	MegaespheraPhylotipo1	0.6690	20	HSV12	0.0768
4	GardnerellaVaginallis	0.5994	21	CrispatusCq	0.0766
5	BVNumero	0.5369	22	UU	0.0685
6	MH	0.5206	23	HRHPVCOMPLETE	0.0646
7	Crispatus	0.2398	24	MG	0.0633
8	CT	0.2009	25	EDAD30	0.0534
9	EDADENA	0.1558	26	LRHPVCOMPLETE	0.0513
10	Jensenii	0.1521	27	MultipleHPVCOMPLETE	0.0434
11	UP	0.1427	28	JenseniiCq	0.0389
12	LGasseri	0.1275	29	pHRHPVCOMPLETE	0.0370
13	InersCq	0.1247	30	LIners	0.0278
14	ELSY	0.1195	31	GasseriCq	0.0234
15	CitologiaBICAT	0.1140	32	MY0911_A	0.0221
16	Citolog	0.0883	33	RMY0911ELSY	0.0196
17	CitologiaOrd	0.0883	34	NG	0.0000

Con base en el ranking obtenido, se contemplan como altamente relevantes atributos como “Atopobium”, “BVCombination”, “MegaespheraPhylotipo1”, “GardnerellaVaginallis”, “BVNumero”, “MH”, “Crispatus”, “CT”. Los atributos anteriormente mencionados obtuvieron un VRM por encima de la media (0.019).

4.2.10 Correlación de Spearman

La identificación de los atributos más relevantes de la VB mediante el método *Correlación de Spearman* se efectuó siguiendo el proceso detallado en Sección 2.4. Es decir, se realizaron 30 corridas del método bajo un esquema de validación cruzada de 10 pliegues. Como resultado, se obtuvieron 30 coeficientes de correlación con base en el método de Spearman para cada atributo del conjunto de datos, los cuales se consideran como el valor de relevancia media. A su vez, estos fueron promediados para

obtener un valor de relevancia medio (VRM). Este valor representa el nivel de relevancia de los atributos para la VB. Los VRM resultantes para cada atributo se representan en la Figura 16.

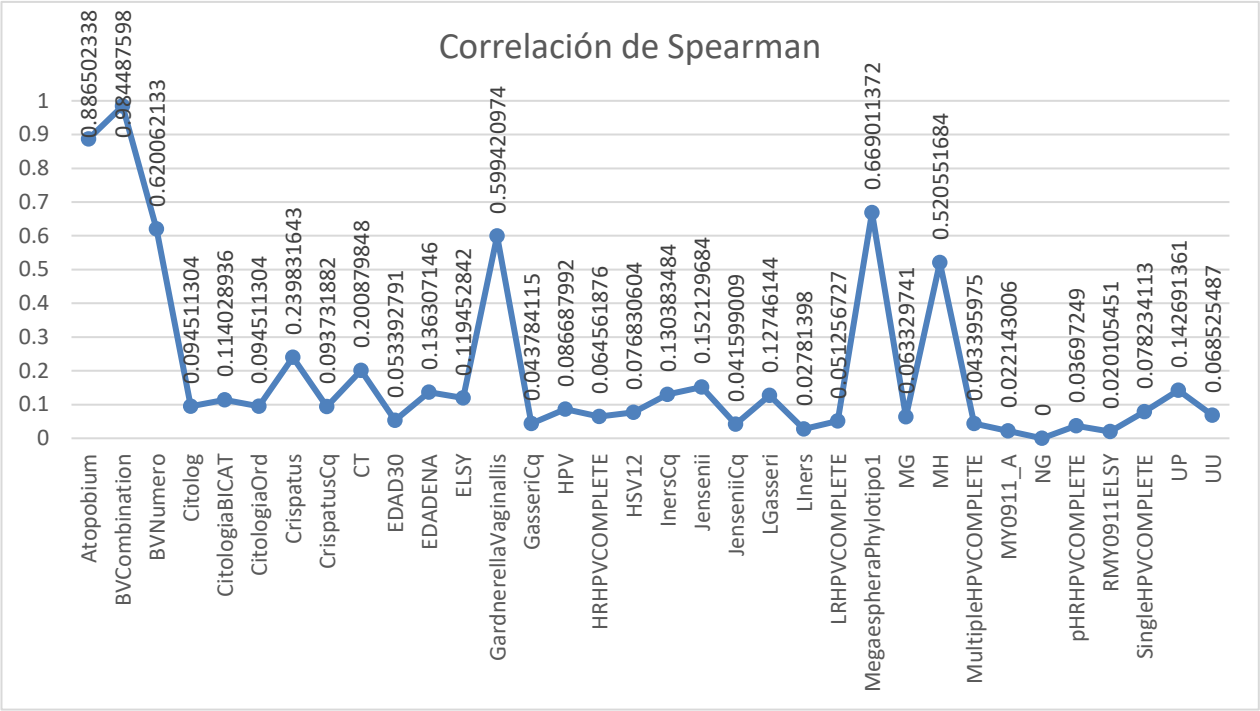

Figura 16. Valor de relevancia media (VRM) obtenida por los atributos de la vaginosis bacteriana en las 30 corridas del método *Correlación de Spearman.*

Con base en los resultados obtenidos, se crea el ranking individual de atributos conforme el criterio de *correlación* de *Spearman.* Para ello, los atributos en el conjunto de datos de VB se ordenaron de forma descendente al considerar el VRM obtenida de las 30 corridas. El ranking individual de atributos de la vaginosis bacteriana basado en el método *correlación de Spearman* se muestra en la Tabla 14.

Tabla 14. Ranking individual de atributos de vaginosis bacteriana generado po0r el método *Correlación de Spearman* mediante las 30 corridas bajo esquema de validación cruzada. Los atributos se ordenan de forma descendente de acuerdo al valor de relevancia media (VRM).

#Rank	Atributo	VRM	#Rank	Atributo	VRM
1	BVCombination	0.9845	18	CrispatusCq	0.0937
2	Atopobium	0.8865	19	HPV	0.0867
3	MegaespheraPhylotipo1	0.6690	20	SingleHPVCOMPLETE	0.0782
4	BVNumero	0.6201	21	HSV12	0.0768
5	GardnerellaVaginallis	0.5994	22	UU	0.0685
6	MH	0.5206	23	HRHPVCOMPLETE	0.0646
7	Crispatus	0.2398	24	MG	0.0633
8	CT	0.2009	25	EDAD30	0.0534
9	Jensenii	0.1521	26	LRHPVCOMPLETE	0.0513
10	UP	0.1427	27	GasseriCq	0.0438
11	EDADENA	0.1363	28	MultipleHPVCOMPLETE	0.0434
12	InersCq	0.1304	29	JenseniiCq	0.0416
13	LGasseri	0.1275	30	pHRHPVCOMPLETE	0.0370
14	ELSY	0.1195	31	LIners	0.0278
15	CitologiaBICAT	0.1140	32	MY0911_A	0.0221
16	Citolog	0.0945	33	RMY0911ELSY	0.0201
17	CitologiaOrd	0.0945	34	NG	0.0000

Respecto a los atributos identificados como relevantes mediante el método de Spearman, ocho de ellos tienen VRM por encima de la media aritmética (0.1971): “BVCombination”, “Atopobium”, “MegaespheraPhylotipo1”, “BVNumero”, “GardnerellaVaginallis”, “MH”, “Crispatus”, “CT”. En ese mismo orden son los atributos de mayor relevancia para la VB en concordancia con este método.

4.2.11 Boruta

La identificación de los atributos más relevantes para el diagnóstico de la vaginosis bacteriana mediante el algoritmo *Boruta* fueron implementados de la siguiente manera. Se realizaron 30 corridas del método bajo un esquema de validación cruzada de 10 pliegues. En cada iteración del esquema de validación cruzada se utilizó el 90% de las instancias para crear un conjunto de entrenamiento, la cual es aleatoriamente cambiado en cada iteración. Por cada una de las iteraciones, se calculó en nivel de relevancia de cada atributo del conjunto de entrenamiento basado en el criterio de ranqueo del algoritmo *Boruta* descrito en la sección 2.4. En la Figura 17 se muestra el resultado de

una corrida del método selector de atributos. En cada corrida, se obtuvo un valor de relevancia para cada atributo al promediar las 10 medidas obtenidas en las iteraciones. Cada corrida fue implementada con semillas diferentes para la aleatoriedad de los datos. Al final, las 30 medidas de relevancia fueron promediadas, con lo cual se obtuvo un valor de relevancia medio (VRM). Este valor es utilizado como criterio de ranqueo.

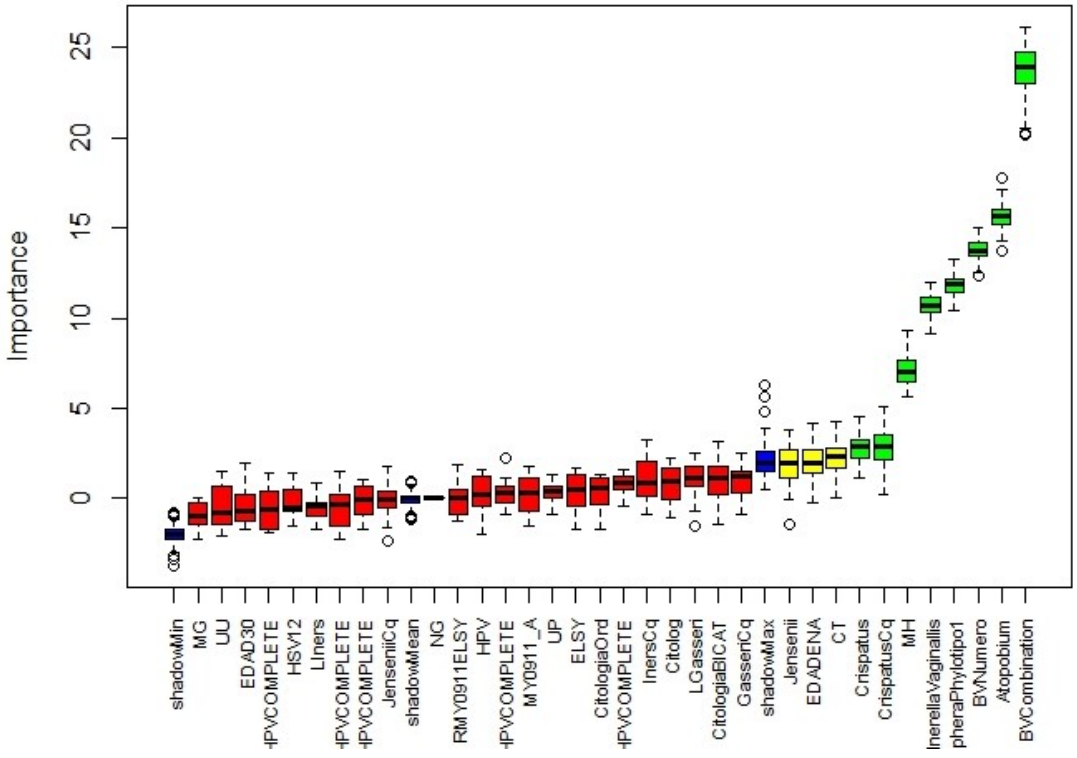

Figura 17. Muestra de una corrida del nivel de relevancia de los atributos de la vaginosis bacteriana calculados mediante el método de selección de atributos *Boruta.* Las cajas azules corresponden al mínimo, promedio y máximos *Z-score* de un atributo sombra. Las cajas rojas y verdes representan las *Z-score* de los atributos rechazados y confirmados, respectivamente

Nótese que las cajas rojas tienen un *Z-score* más bajo que las *Z-score* máxima de los atributos sombra. Es por ello que dichos atributos se colocaron en una categoría “sin importancia”. De manera contraria, las cajas azules tienen un *Z-score* más alto que los atributos sombra, por tanto, fueron catalogadas como “relevantes”. En la Figura 18 se muestran los VRM obtenidos por los atributos con base en las 30 corridas de *Boruta*.

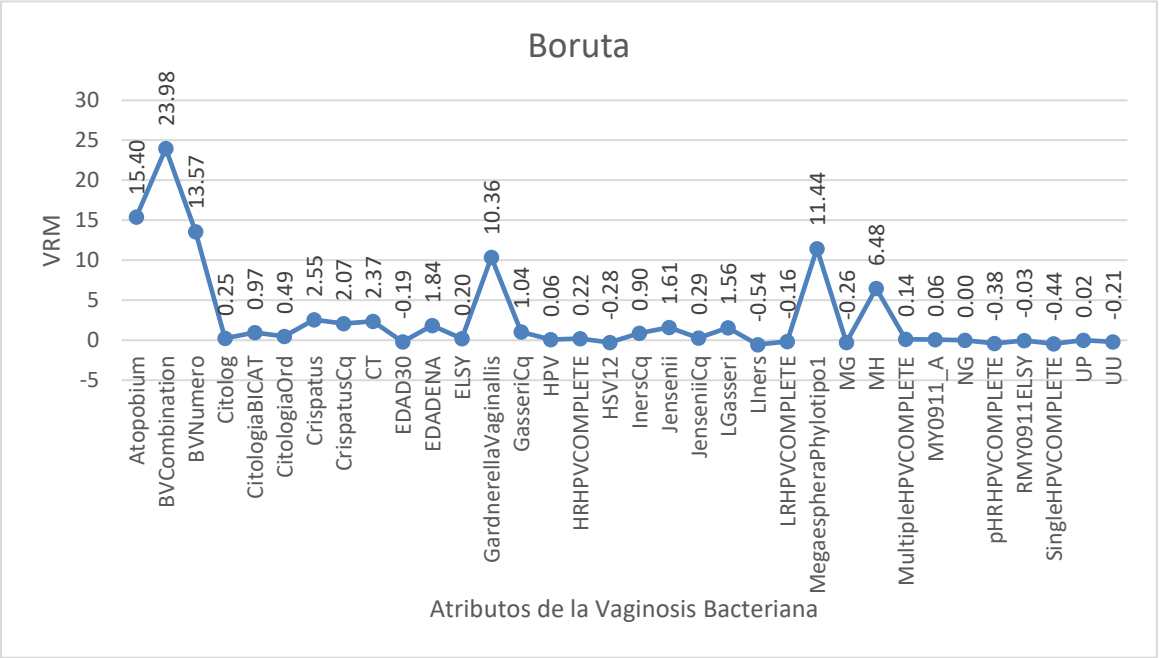

Figura 18. Valores de relevancia media (VRM) obtenida por los atributos del conjunto de vaginosis bacteriana mediante las 30 ejecuciones con validación cruzada del método de selección de atributos *Boruta*.

Con base en estos resultados, se crea el ranking individual de atributos. Para ello, los atributos en el conjunto de datos son ordenados de forma ascendente respecto al VRM obtenido de las 30 corridas. El ranking individual de atributos de la vaginosis bacteriana basado en el método *Boruta* es mostrado en la Tabla 15.

Tabla 15. Ranking individual de atributos de vaginosis bacteriana generado mediante el método *Boruta*. Los atributos se ordenan de forma descendente de acuerdo al Valor de Relevancia Medio (VRM) -promedio de los 30 resultados-.

#Rank	Atributos	VRM	#Rank	Atributos	VRM
1	BVCombination	23.9754183	18	Citolog	0.25150454
2	Atopobium	15.4040995	19	HRHPVCOMPLETE	0.22155207
3	BVNumero	13.5673566	20	ELSY	0.20344599
4	MegaespheraPhylotipo1	11.4388616	21	MultipleHPVCOMPLETE	0.13560746
5	GardnerellaVaginallis	10.3636341	22	MY0911_A	0.06442591
6	MH	6.47831605	23	HPV	0.06375545
7	Crispatus	2.55088224	24	UP	0.01654041
8	CT	2.36929792	25	NG	0
9	CrispatusCq	2.07247704	26	RMY0911ELSY	-0.03499371
10	EDADENA	1.84023119	27	LRHPVCOMPLETE	-0.15526681
11	Jensenii	1.6133249	28	EDAD30	-0.18693299
12	LGasseri	1.5606291	29	UU	-0.20967706
13	GasseriCq	1.04037678	30	MG	-0.26116644
14	CitologiaBICAT	0.972775	31	HSV12	-0.27675027
15	InersCq	0.89758364	32	pHRHPVCOMPLETE	-0.38424318
16	CitologiaOrd	0.48637596	33	SingleHPVCOMPLETE	-0.43565114
17	JenseniiCq	0.2853279	34	LIners	-0.54292161

Los resultados muestran que atributos como “BVCombination”, “Atopobium”, “BVNumero”, “MegaespheraPhylotipo1”, “GardnerellaVaginallis” y “MH” obtuvieron VRM por encima de la media aritmética (2.8054), lo que los identifica como los más relevantes para la VB.

4.2.12 Consistencia

La identificación de los atributos más relevantes de la VB mediante el algoritmo *consistencia* se efectuó siguiendo el proceso descrito en Sección 2.4. Esto es, se realizaron 30 corridas del método bajo un esquema de validación cruzada de 10 pliegues. Por cada iteración de la validación cruzada se obtuvo un subconjunto de datos como resultado de la implementación de *consistencia*. Este subconjunto representa los atributos identificados como más relevantes de la VB respecto al método en cuestión. Al final, se obtuvieron 300 subconjuntos de datos. A partir de estos subconjuntos obtenidos, se realizó un análisis de distribución de frecuencias para identificar el número de veces que un atributo se identifica entre los subconjuntos de atributos creados mediante *consistencia*. Los resultados del análisis de frecuencia se representan en la Figura 19.

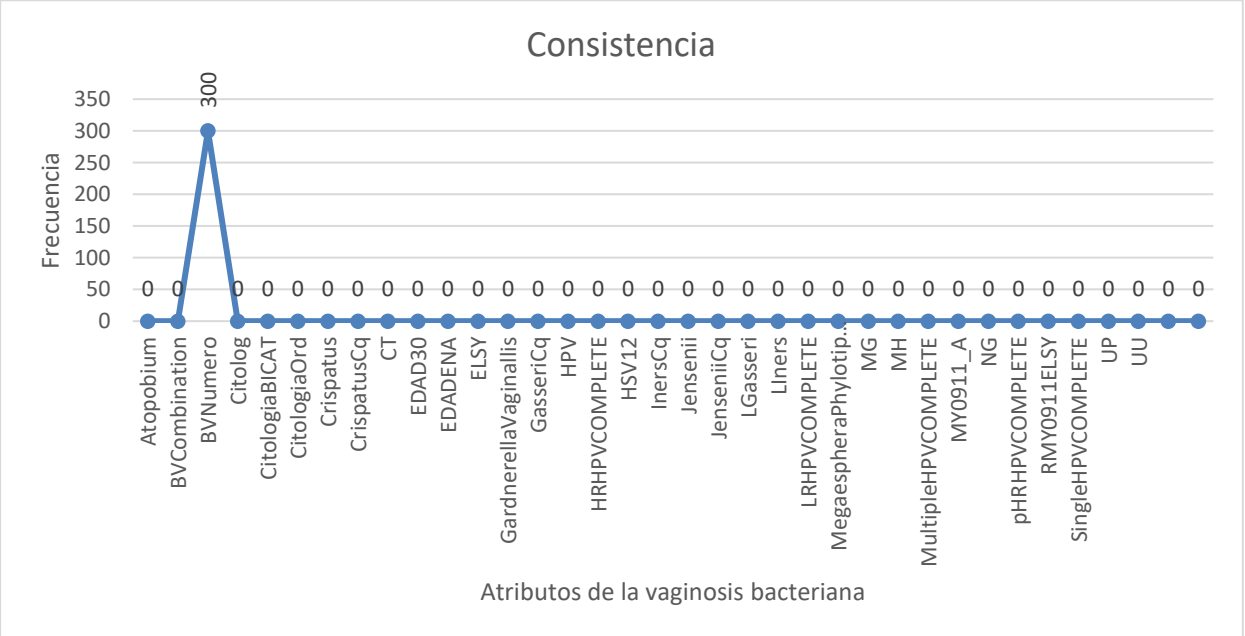

Figura 19. Distribución de frecuencias de los atributos de la vaginosis bacteriana a través de las 30 corridas con VC-10 del método selección *consistencia*.

Con base en los resultados obtenidos, se crea el ranking individual de atributos mediante el método *consistencia.* Para ello, los atributos en el conjunto de datos de VB son ordenados de forma descendente al considerar la frecuencia obtenida de las 30 corridas (300 resultados) como criterio de ordenamiento. El ranking individual de atributos de la vaginosis bacteriana basado en el método *consistencia* se muestra en la Tabla 16.

Tabla 16. Ranking individual de atributos de vaginosis bacteriana generado por el algoritmo *consistencia* mediante las 30 corridas bajo esquema de validación cruzada. Los atributos se ordenan de forma descendente de acuerdo a la frecuencia obtenida de los resultados.

#Rank	Atributo	Frecuencia	#Rank	Atributo	Frecuencia
1	BVNumero	300	18	InersCq	0
2	Atopobium	0	19	Jensenii	0
3	BVCombination	0	20	JenseniiCq	0
4	Citolog	0	21	LGasseri	0
5	CitologiaBICAT	0	22	LIners	0
6	CitologiaOrd	0	23	LRHPVCOMPLETE	0
7	Crispatus	0	24	MegaespheraPhylotipo1	0
8	CrispatusCq	0	25	MG	0
9	CT	0	26	MH	0
10	EDAD30	0	27	MultipleHPVCOMPLETE	0
11	EDADENA	0	28	MY0911_A	0
12	ELSY	0	29	NG	0
13	GardnerellaVaginallis	0	30	pHRHPVCOMPLETE	0
14	GasseriCq	0	31	RMY0911ELSY	0
15	HPV	0	32	SingleHPVCOMPLETE	0
16	HRHPVCOMPLETE	0	33	UP	0
17	HSV12	0	34	UU	0

Respecto a los resultados obtenidos por la corrida del método *consistencia*, solo uno de 34 atributos que conforman el conjunto de datos fue identificado como relevante para la vaginosis bacteriana: “BVNumero”.

4.2.13 Selección secuencial hacia adelante (SFS)

La identificación de los atributos más relevantes para el diagnóstico de la vaginosis bacteriana mediante el algoritmo selección secuencias hacia adelante (del inglés *Sequential Forward Selection (SFS))* se efectuó de la siguiente manera. Se realizaron

30 corridas del método bajo un esquema de validación cruzada de 10 pliegues. Cada corrida fue implementada con semillas diferentes para asegurar la aleatoriedad de los datos. En cada iteración del esquema de validación cruzada se utilizó el 90% de las instancias para crear un conjunto de entrenamiento, la cual es aleatoriamente cambiado en cada iteración. En cada una de las iteraciones se implementó el método *SFS* sobre el conjunto de entrenamiento. Por cada iteración se obtuvo un subconjunto de datos como resultado de la implementación del *SFS*. Este subconjunto representa los atributos identificados como más relevantes de la VB respecto al método en cuestión. Al finalizar, se obtuvieron 300 subconjuntos de datos. A partir de estos subconjuntos, se realizó un análisis de distribución de frecuencias, que permite identificar el número de veces que un atributo se encuentra entre los subconjuntos de atributos resultantes con el método *SFS*. Los resultados del análisis de frecuencia se representan en la Figura 20.

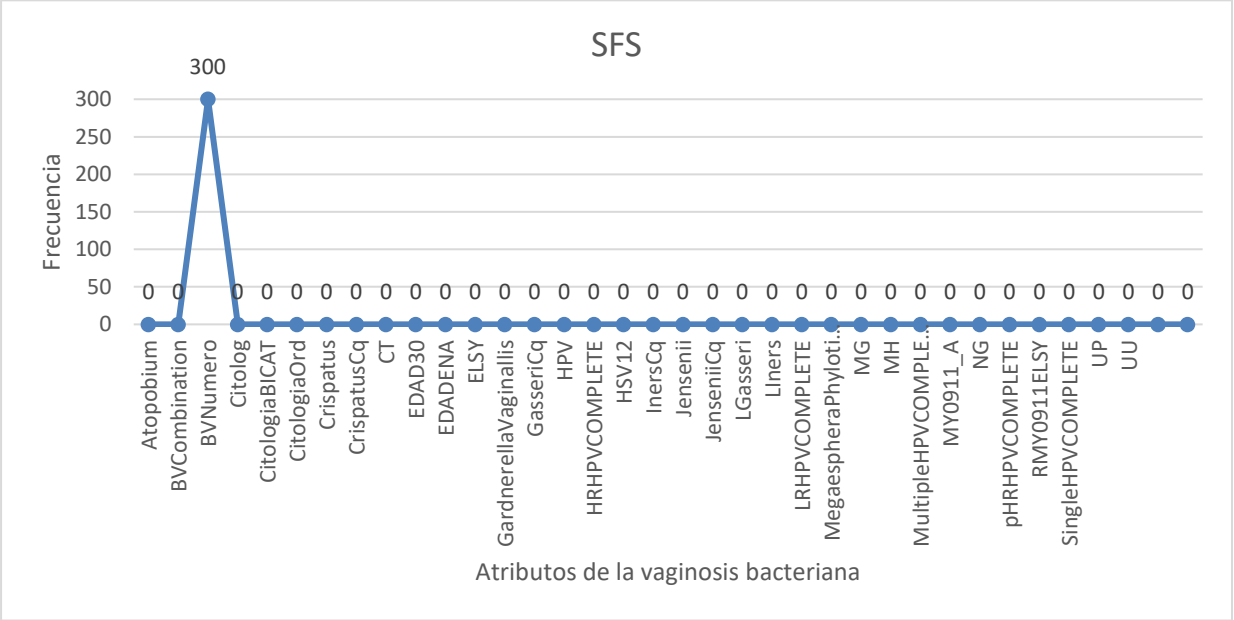

Figura 20. Distribución de frecuencias de los atributos de la vaginosis bacteriana a través de las 30 corridas con VC-10 del método S*FS*.

Con base en los resultados obtenidos, se crea el ranking individual de atributos mediante el método *SFS.* Para ello, los atributos en el conjunto de datos de VB son ordenados de forma descendente al considerar la frecuencia obtenida de las 30 corridas como criterio de ordenamiento. El ranking individual de atributos de la vaginosis bacteriana basado en el método S*FS* se muestra en la Tabla 17.

Tabla 17. Ranking individual de atributos de vaginosis bacteriana generado por el algoritmo *selección secuencial hacia adelante* mediante las 30 corridas bajo esquema de validación cruzada. Los atributos se ordenan de forma descendente de acuerdo a la frecuencia obtenida de los 300 resultados.

#Rank	Atributo	Frecuencia	#Rank	Atributo	Frecuencia
1	BVNumero	300	18	InersCq	0
2	Atopobium	0	19	Jensenii	0
3	BVCombination	0	20	JenseniiCq	0
4	Citolog	0	21	LGasseri	0
5	CitologiaBICAT	0	22	LIners	0
6	CitologiaOrd	0	23	LRHPVCOMPLETE	0
7	Crispatus	0	24	MegaespheraPhylotipo1	0
8	CrispatusCq	0	25	MG	0
9	CT	0	26	MH	0
10	EDAD30	0	27	MultipleHPVCOMPLETE	0
11	EDADENA	0	28	MY0911_A	0
12	ELSY	0	29	NG	0
13	GardnerellaVaginallis	0	30	pHRHPVCOMPLETE	0
14	GasseriCq	0	31	RMY0911ELSY	0
15	HPV	0	32	SingleHPVCOMPLETE	0
16	HRHPVCOMPLETE	0	33	UP	0
17	HSV12	0	34	UU	0

A través de las 30 corridas con VC-10, el método *SFS* identificó como único atributo relevante a “BVNumero”. Los demás atributos se identificaron como irrelevantes para el conjunto de datos de VB.

4.2.14 Selección secuencial hacia atrás (SBS)

La identificación de los atributos más relevantes para el diagnóstico de la vaginosis bacteriana mediante el algoritmo s elección secuencial hacia atrás (del inglés *sequential backward selection, SBS)* se efectuó siguiendo el proceso descrito en Sección 2.4. Se realizaron 30 corridas del método bajo un esquema de validación cruzada de 10 pliegues. Por cada iteración se obtuvo un subconjunto de datos como resultado de la implementación del *SBS*. Este subconjunto representa los atributos identificados como más relevantes de la VB respecto al método en cuestión. Al finalizar, se obtuvieron 300 subconjuntos de datos. A partir de estos subconjuntos, se realizó un análisis de distribución de frecuencias, que permite identificar el número de veces que un atributo se identifica entre los subconjuntos de atributos resultantes con el método SBS. Los resultados del análisis de frecuencia se representan en la Figura 21.

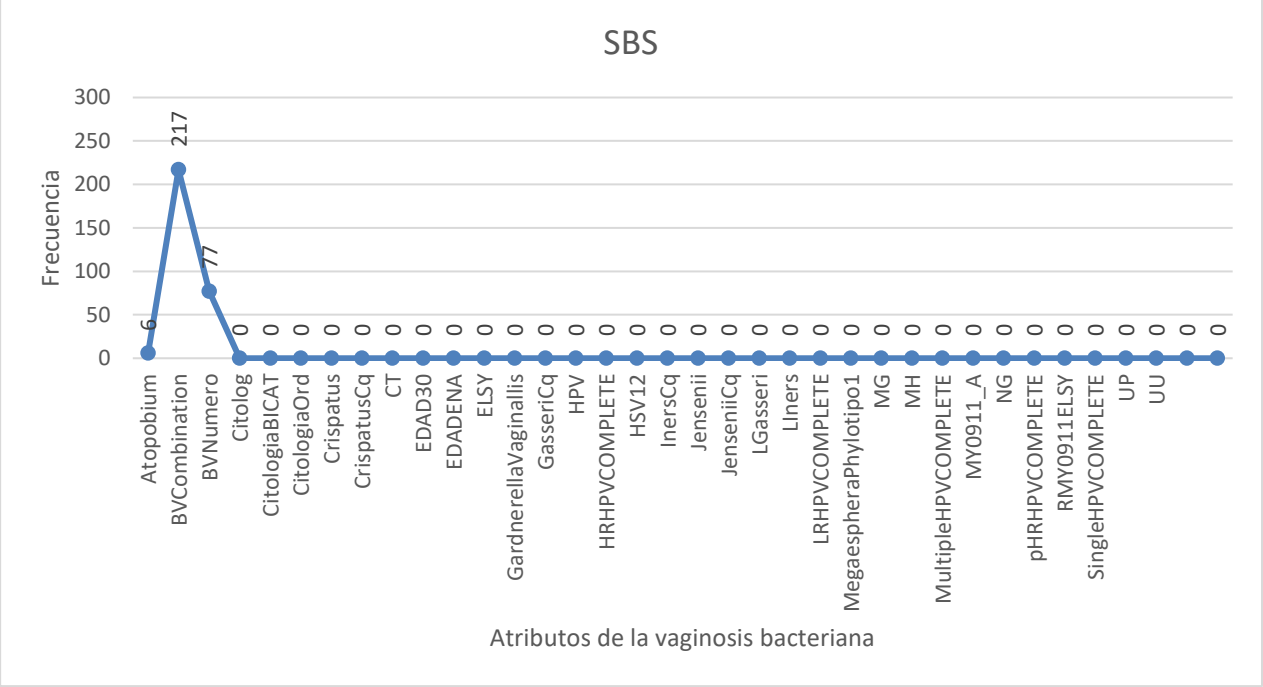

Figura 21. Distribución de frecuencias de los atributos de la vaginosis bacteriana a través de las 30 corridas con VC-10 del método *selección secuencial hacia atrás* -del inglés *sequential backward selection* (*SBS)-.*

.

Con base en los resultados obtenidos, se crea el ranking individual de atributos mediante el método *SBS.* Para ello, los atributos en el conjunto de datos de VB son ordenados de forma descendente al considerar la frecuencia obtenida de las 30 corridas como criterio de ordenamiento. El ranking individual de atributos de la vaginosis bacteriana basado en el método S*BS* se muestra en la Tabla 18.

Tabla 18. Ranking individual de atributos de vaginosis bacteriana generado mediante el algoritmo selección secuencial hacia atrás -del inglés *sequential backward selection (SBS)*- mediante las 30 corridas bajo esquema de validación cruzada. Los atributos se ordenan de forma descendente de acuerdo a la frecuencia obtenida de los 300 resultados.

#Rank	Atributo	Frecuencia	#Rank	Atributo	Frecuencia
1	BVCombination	217	18	InersCq	0
2	BVNumero	77	19	Jensenii	0
3	Atopobium	6	20	JenseniiCq	0
4	Citolog	0	21	LGasseri	0
5	CitologiaBICAT	0	22	LIners	0
6	CitologiaOrd	0	23	LRHPVCOMPLETE	0
7	Crispatus	0	24	MegaespheraPhylotipo1	0
8	CrispatusCq	0	25	MG	0
9	CT	0	26	MH	0
10	EDAD30	0	27	MultipleHPVCOMPLETE	0
11	EDADENA	0	28	MY0911_A	0
12	ELSY	0	29	NG	0
13	GardnerellaVaginallis	0	30	pHRHPVCOMPLETE	0
14	GasseriCq	0	31	RMY0911ELSY	0
15	HPV	0	32	SingleHPVCOMPLETE	0
16	HRHPVCOMPLETE	0	33	UP	0
17	HSV12	0	34	UU	0

De forma particular, SBS identificó como relevantes tres de los 34 atributos que conforman los atributos de la VB. “BVCombination”, “BVNumero” y “Atopobium” son los atributos más importantes respecto al método SBS.

4.2.15 Selección secuencial flotante hacia adelante (SFFS)

La identificación de los atributos más relevantes de la VB mediante el algoritmo selección secuencial flotante hacia adelante *-sequential forward floating selection (SFFS)-* se efectuó siguiendo el proceso descrito en Sección 2.4. Se realizaron 30 corridas del método bajo un esquema de validación cruzada de 10 pliegues. Por cada iteración de la validación cruzada se obtuvo un subconjunto de datos como resultado de la implementación del *SFFS*. Este subconjunto representa los atributos identificados como más relevantes de la VB respecto al método en cuestión. Al final, se obtuvieron 300 subconjuntos de datos. A partir de estos subconjuntos obtenidos, se realizó un análisis de distribución de frecuencias para identificar el número de veces que un atributo se identifica entre los subconjuntos de atributos creados mediante *SFFS*. Los resultados del análisis de frecuencia se representan en la Figura 22.

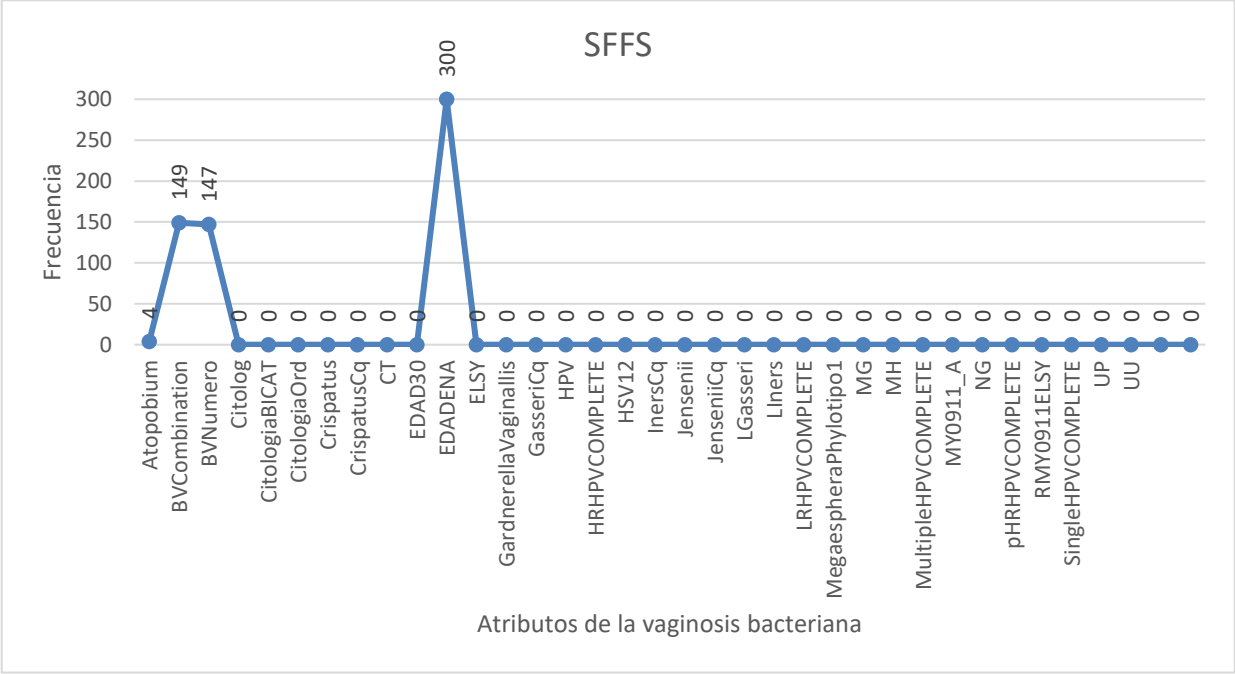

Figura 22. Distribución de frecuencias de los atributos de la vaginosis bacteriana a través de las 30 corridas con VC-10 del método selección secuencial flotante hacia adelante.

Con base en los resultados obtenidos, se crea el ranking individual de atributos mediante el método *SFFS.* Para ello, los atributos en el conjunto de datos de VB se ordenaron de forma descendente al considerar la frecuencia obtenida de las 30 corridas (300 resultados) como criterio de ordenamiento. El ranking individual de atributos de la vaginosis bacteriana basado en el método S*FFS* se muestra en la Tabla 19.

Tabla 19. Ranking individual de atributos de vaginosis bacteriana generado por el algoritmo selección secuencial flotante hacia adelante -*sequential fordward floating selection* (*SFFS*)- mediante las 30 corridas bajo esquema de validación cruzada. Los atributos se ordenan de forma descendente de acuerdo a la frecuencia obtenida de los 300 subconjuntos de atributos resultantes.

#Rank	Atributo	Frecuencia	#Rank	Atributo	Frecuencia
1	EDADENA	300	18	InersCq	0
2	BVCombination	149	19	Jensenii	0
3	BVNumero	147	20	JenseniiCq	0
4	Atopobium	4	21	LGasseri	0
5	Citolog	0	22	LIners	0
6	CitologiaBICAT	0	23	LRHPVCOMPLETE	0
7	CitologiaOrd	0	24	MegaespheraPhylotipo1	0
8	Crispatus	0	25	MG	0
9	CrispatusCq	0	26	MH	0
10	CT	0	27	MultipleHPVCOMPLETE	0
11	EDAD30	0	28	MY0911_A	0
12	ELSY	0	29	NG	0
13	GardnerellaVaginallis	0	30	pHRHPVCOMPLETE	0
14	GasseriCq	0	31	RMY0911ELSY	0
15	HPV	0	32	SingleHPVCOMPLETE	0
16	HRHPVCOMPLETE	0	33	UP	0
17	HSV12	0	34	UU	0

Particularmente, *SFFS* identificó como relevantes solamente cuatro atributos entre las características de la VB: “EDADENA”, “BVCombination”, “BVNumero” y “Atopobium”. Los demás, es decir los restantes 30 atributos, fueron catalogados como no relevantes.

4.2.16 Selección secuencial flotante hacia atrás (SBFS)

La identificación de los atributos más relevantes de la VB mediante el algoritmo selección secuencial flotante hacia atrás -*sequential backward floating selection* (*SBFS)*- se efectuó siguiendo el proceso descrito en Sección 2.4. Se realizaron 30 corridas del método bajo un esquema de validación cruzada de 10 pliegues. Por cada iteración se obtuvo un subconjunto de datos como resultado de la implementación del *SBFS*. Este subconjunto representa los atributos identificados como más relevantes de la VB respecto al método en cuestión. Al final, se obtuvieron 300 subconjuntos de datos. A partir de estos subconjuntos obtenidos, se realizó un análisis de distribución de frecuencias para identificar el número de veces que un atributo se identifica entre los subconjuntos de atributos creados mediante *SBFS*. Los resultados del análisis de frecuencia se representan en la Figura 23.

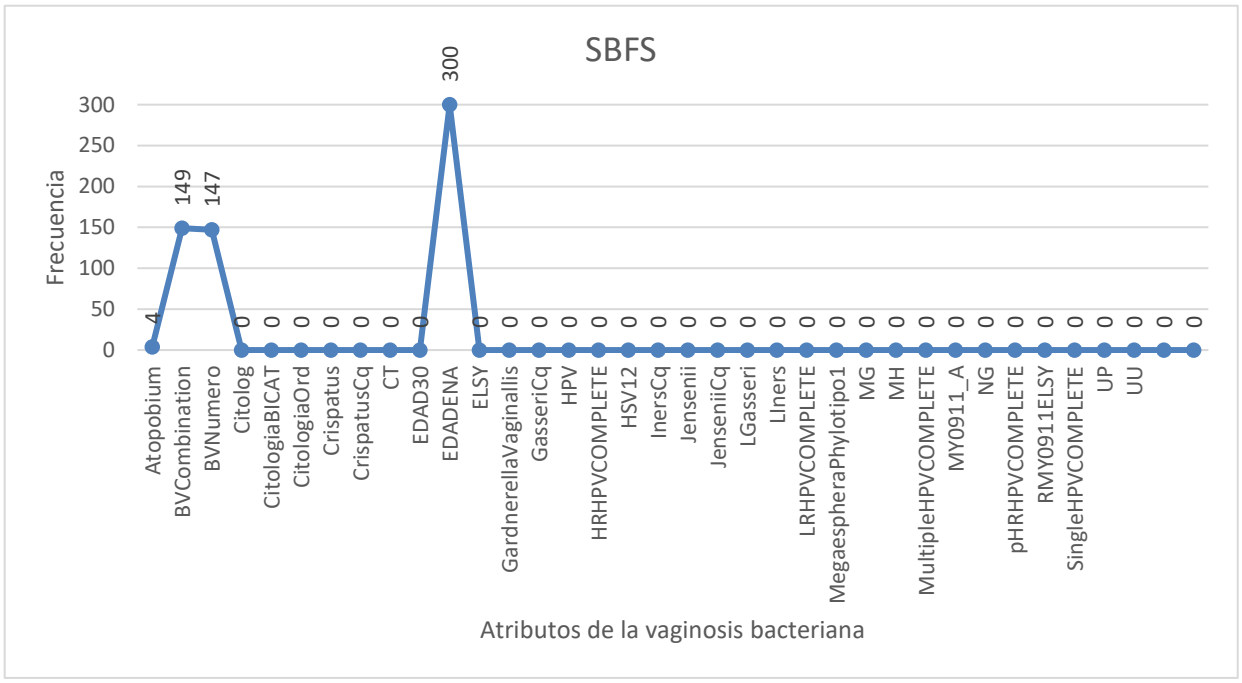

Figura 23. Distribución de frecuencias de los atributos de la vaginosis bacteriana a través de las 30 corridas con VC-10 del método selección secuencial flotante hacia atrás.

Con base en los resultados obtenidos, se crea el ranking individual de atributos mediante el método *SBFS.* Para ello, los atributos en el conjunto de datos de VB son ordenados de forma descendente al considerar la frecuencia obtenida de las 30 corridas (300 resultados) como criterio de ordenamiento. El ranking individual de atributos de la vaginosis bacteriana basado en el método S*BFS* se muestra en la Tabla 20.

Tabla 20. Ranking individual de atributos de vaginosis bacteriana generado por el algoritmo *selección secuencial flotante hacia atrás* mediante las 30 corridas bajo esquema de validación cruzada. Los atributos se ordenan de forma descendente de acuerdo a la frecuencia obtenida de los 300 subconjuntos de atributos resultantes.

#Rank	Atributo	Frecuencia	#Rank	Atributo	Frecuencia
1	EDADENA	300	18	InersCq	0
2	BVCombination	149	19	Jensenii	0
3	BVNumero	147	20	JenseniiCq	0
4	Atopobium	4	21	LGasseri	0
5	Citolog	0	22	LIners	0
6	CitologiaBICAT	0	23	LRHPVCOMPLETE	0
7	CitologiaOrd	0	24	MegaespheraPhylotipo1	0
8	Crispatus	0	25	MG	0
9	CrispatusCq	0	26	MH	0
10	CT	0	27	MultipleHPVCOMPLETE	0
11	EDAD30	0	28	MY0911_A	0
12	ELSY	0	29	NG	0
13	GardnerellaVaginallis	0	30	pHRHPVCOMPLETE	0
14	GasseriCq	0	31	RMY0911ELSY	0
15	HPV	0	32	SingleHPVCOMPLETE	0
16	HRHPVCOMPLETE	0	33	UP	0
17	HSV12	0	34	UU	0

En este caso, el *SBFS* identificó cuatro atributos relevantes entre las características de la VB: “EDADENA”, “BVCombination”, “BVNumero” y “Atopobium”. Esto significa, que para el método *SBFS*, 30 atributos se consideraron como irrelevantes para la enfermedad en cuestión.

4.2.17 Peso de los atributos (Maquina de vector soporte)

La identificación de los atributos más relevantes para el diagnóstico de la vaginosis bacteriana mediante el algoritmo *SVM* fue implementado de la siguiente manera. Se realizaron 30 corridas del método bajo un esquema de validación cruzada de 10 pliegues. En cada iteración del esquema de validación cruzada se utilizó el 90% de las instancias para crear un conjunto de entrenamiento, la cual es aleatoriamente cambiado en cada iteración. Por cada una de las iteraciones, se calculó en nivel de relevancia de cada atributo del conjunto de entrenamiento basado en el peso de los atributos con *SVM* tal como se describió anteriormente. En cada corrida, se obtuvo un valor de relevancia para cada atributo al promediar las 10 medidas obtenidas en las iteraciones. Al final, las 30 medidas de relevancia se promediaron, con lo cual se obtuvo un valor de relevancia medio (VRM). Cada corrida fue implementada con semillas diferentes para asegurar la aleatoriedad de los datos. En la Figura 24 se muestran los VRM resultantes de cada atributo obtenido a través del cálculo de los pesos de los atributos mediante *SVM.*

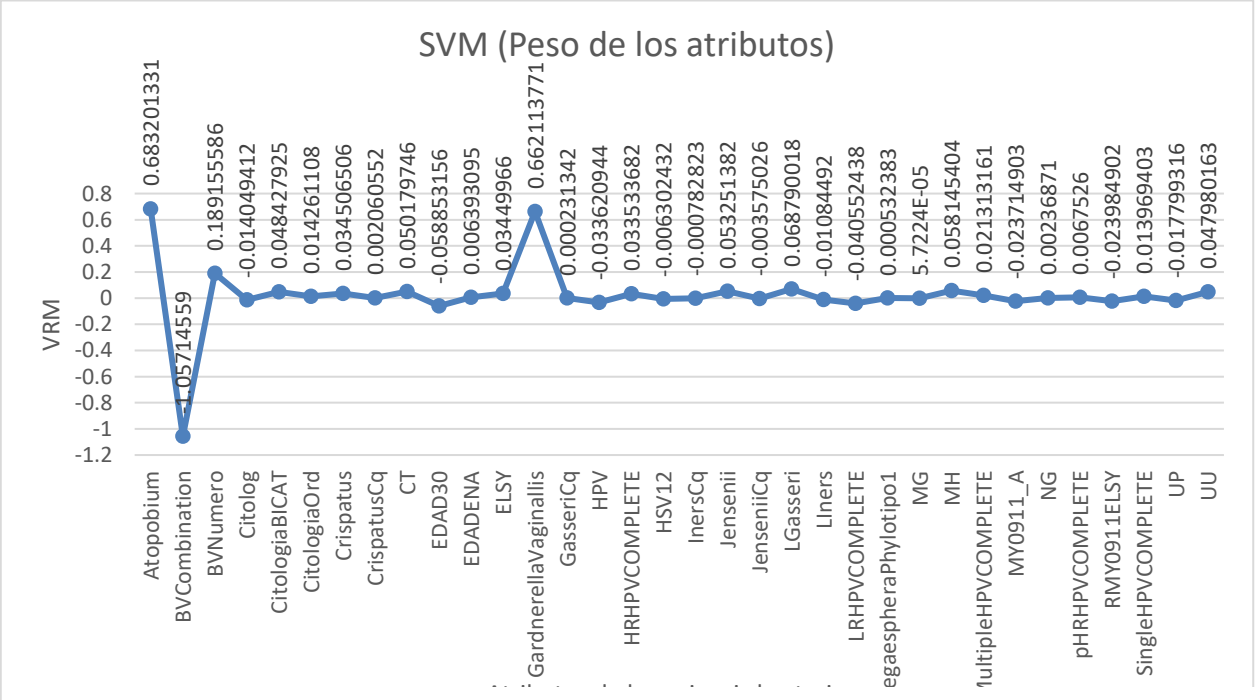

Figura 24. Valor de relevancia media (VRM) obtenida por los atributos de la vaginosis bacteriana en las 30 corridas del método *peso de los atributos mediante SVM*

Con base en estos resultados, se crea el ranking individual de atributos mediante el método *SVM feature weights*. Para ello, los atributos en el conjunto de datos son ordenados de forma descendente respecto al VRM obtenido de las 30 corridas. El ranking individual de atributos de la vaginosis bacteriana basado en el método *SVM feature weights* se muestra en la Tabla 21.

Tabla 21. Ranking individual de atributos de vaginosis bacteriana generado mediante el método *SVM feature weights*. Los atributos se ordenan de forma descendente de acuerdo al Valor de Relevancia Medio (VRM) -promedio de los 30 resultados-.

#Rank	Atributos	VRM	#Rank	Atributos	VRM
1	Atopobium	0.68320133	18	NG	0.0023687
2	GardnerellaVaginallis	0.66211377	19	CrispatusCq	0.002060552
3	BVNumero	0.18915559	20	MegaespheraPhylotipo1	0.000532383
4	LGasseri	0.06879002	21	GasseriCq	0.000231342
5	MH	0.0581454	22	MG	5.7224E-05
6	Jensenii	0.05325138	23	InersCq	-0.00078282
7	CT	0.05017975	24	JenseniiCq	-0.00357503
8	CitologiaBICAT	0.04842792	25	HSV12	-0.00630243
9	UU	0.04798016	26	LIners	-0.01084492
10	Crispatus	0.03450651	27	Citolog	-0.01404941
11	ELSY	0.03449966	28	UP	-0.01779932
12	HRHPVCOMPLETE	0.03353368	29	MY0911_A	-0.0237149
13	MultipleHPVCOMPLETE	0.02131316	30	RMY0911ELSY	-0.0239849
14	CitologiaOrd	0.01426111	31	HPV	-0.03362094
15	SingleHPVCOMPLETE	0.0139694	32	LRHPVCOMPLETE	-0.04055244
16	pHRHPVCOMPLETE	0.0067526	33	EDAD30	-0.05885316
17	EDADENA	0.00639309	34	BVCombination	-1.05714559

4.2.18 Coeficiente de correlación por regresión logística

La identificación de los atributos más relevantes para el diagnóstico de la vaginosis bacteriana mediante el algoritmo *regresión logística* fueron implementados de la siguiente manera. Se realizaron 30 corridas del método bajo un esquema de validación cruzada de 10 pliegues. En cada iteración del esquema de validación cruzada se utilizó el 90% de las instancias para crear un conjunto de entrenamiento, la cual es aleatoriamente cambiado en cada iteración. Por cada una de las iteraciones, se calculó en nivel de relevancia de cada atributo del conjunto de entrenamiento basado en el

peso de los atributos con *RL.* En cada corrida, se obtuvo un valor de relevancia para cada atributo de acuerdo a la magnitud de su coeficiente de correlación en todos los conjuntos de datos de las validaciones cruzadas dividido por su desviación estándar. Mas información se detalla en la sección 2.4. Al final, las 30 medidas de relevancia se promediaron, con lo cual se obtuvo un valor de relevancia medio (VRM). Cada corrida fue implementada con semillas diferentes para asegurar la aleatoriedad de los datos. En la Figura 25 se muestran los VRM resultantes de cada atributo obtenido a través del cálculo de los pesos de los atributos mediante *RL.*

Con base en estos resultados, se crea el ranking individual de atributos mediante el método de *regresión logística*. Para ello, los atributos en el conjunto de datos son ordenados de forma descendente al considerar el VRM obtenido de las 30 corridas.

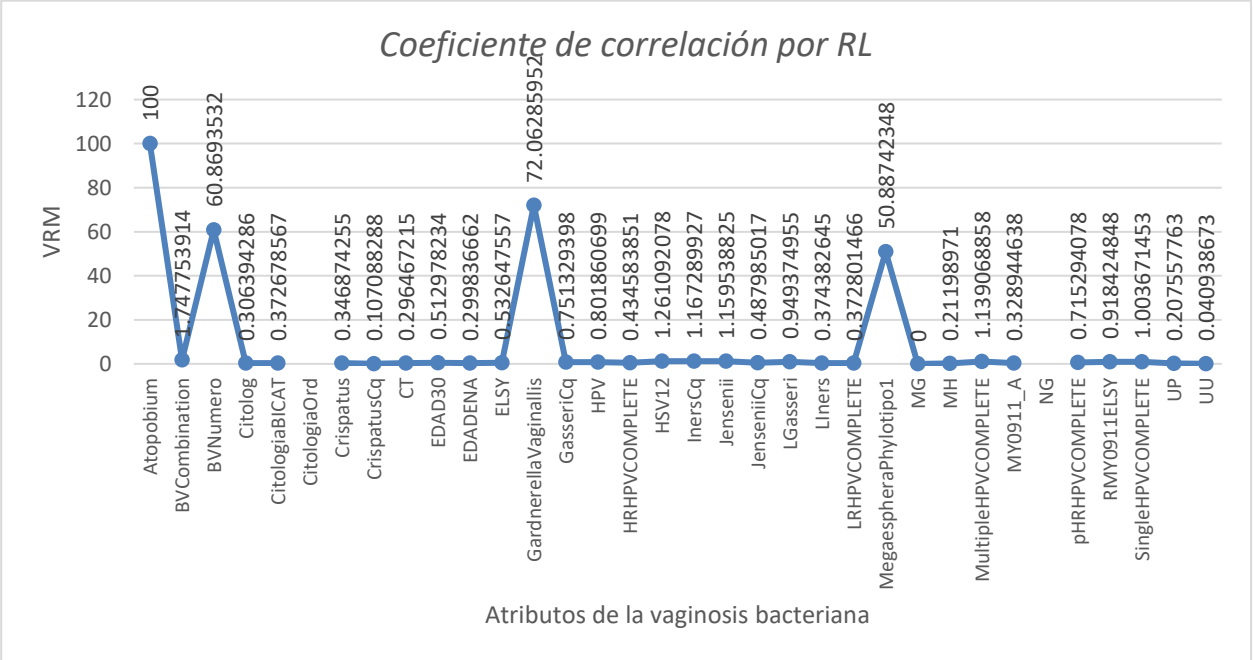

Figura 25. Valor de relevancia media (VRM) obtenida por los atributos de la vaginosis bacteriana en las 30 corridas del método *Regresión logística (RL*).

El ranking individual de atributos de la vaginosis bacteriana basado en el método *RL* se muestra en la Tabla 22.

Tabla 22. Ranking individual de atributos de vaginosis bacteriana generado mediante el método *regresión logística* (RL). Los atributos se ordenan de forma descendente de acuerdo al Valor de Relevancia Medio (VRM) -promedio de los 30 resultados-.

#Rank	Features	VRM	#Rank	Features	VRM
1	Atopobium	100	18	JenseniiCq	0.4879850
2	GardnerellaVaginallis	72.0628595	19	HRHPVCOMPLETE	0.434583851
3	BVNumero	60.8693532	20	LIners	0.374382645
4	MegaespheraPhylotipo1	50.8874235	21	LRHPVCOMPLETE	0.372801466
5	BVCombination	1.74775391	22	CitologiaBICAT	0.372678567
6	HSV12	1.26109208	23	Crispatus	0.346874255
7	InersCq	1.16728993	24	MY0911_A	0.328944638
8	Jensenii	1.15953883	25	Citolog	0.306394286
9	MultipleHPVCOMPLETE	1.13906886	26	EDADENA	0.299836662
10	SingleHPVCOMPLETE	1.00367145	27	CT	0.296467215
11	LGasseri	0.94937496	28	MH	0.21198971
12	RMY0911ELSY	0.91842485	29	UP	0.207557763
13	HPV	0.8018607	30	CrispatusCq	0.107088288
14	GasseriCq	0.7513294	31	UU	0.040938673
15	pHRHPVCOMPLETE	0.71529408	32	CitologiaOrd	0
16	ELSY	0.53264756	33	MG	0
17	EDAD30	0.51297823	34	NG	0

De acuerdo con este método, se identificó el nivel de relevancia para cada atributo en el conjunto de datos de la vaginosis bacteriana, obteniendo lo siguiente. Solamente 4 de 34 atributos de la VB fueron identificados como altamente relevantes al obtener niveles de VRM por encima de la media aritmética (8.843190708). Entre ellos se destacan atributos como “Atopobium”, “GardnerellaVaginallis”, “BVNumero” y “MegaespheraPhylotipo1”.

4.2.19 *OneR*

La implementación de *OneR* como método selector de atributos se describe a continuación. Se realizaron 30 ejecuciones de *OneR* bajo un esquema de validación cruzada de 10 pliegues. En cada iteración de validación cruzada se implementó OneR usando el 90% de las instancias del conjunto de datos de vaginosis bacteriana, es

decir, sobre el conjunto de datos de entrenamiento. Esta porción de instancias de entrenamiento es cambiada de manera aleatoria en cada corrida. Para asegurar la aleatoriedad, se utilizaron semillas diferentes. Al final, los 30 valores de relevancia son promediados, con lo que se obtuvo un valor de relevancia medio (VRM) para cada atributo. En este caso, el conjunto completo de atributos del conjunto de datos fue utilizado, es decir, los 34 atributos.

Los resultados obtenidos mediante la corrida del método *OneR* son presentados. El método *OneR* crea una o unas reglas al basar su modelado en árboles de decisión de un solo nivel, por lo que las reglas resultantes para la clasificación de la vaginosis bacteriana se pueden extraer. Una muestra de las reglas creadas por *OneR* al usar los datos de entrenamiento es la siguiente:

If BVNumero = 0 then VBPCR = 0
If BVNumero = 1 then VBPCR = 1
If BVNumero = 2 then VBPCR = 1
If BVNumero = 3 then VBPCR = 0

Donde VBPCR = 0 representa el diagnóstico de vaginosis bacteriana negativo y VBPCR = 1 como VB positivo.

El nivel de relevancia de los atributos mediante *OneR* fue extraído del modelo de entrenamiento. En la Figura 26 se muestra el promedio de los resultados obtenidos de las 30 ejecuciones del método.

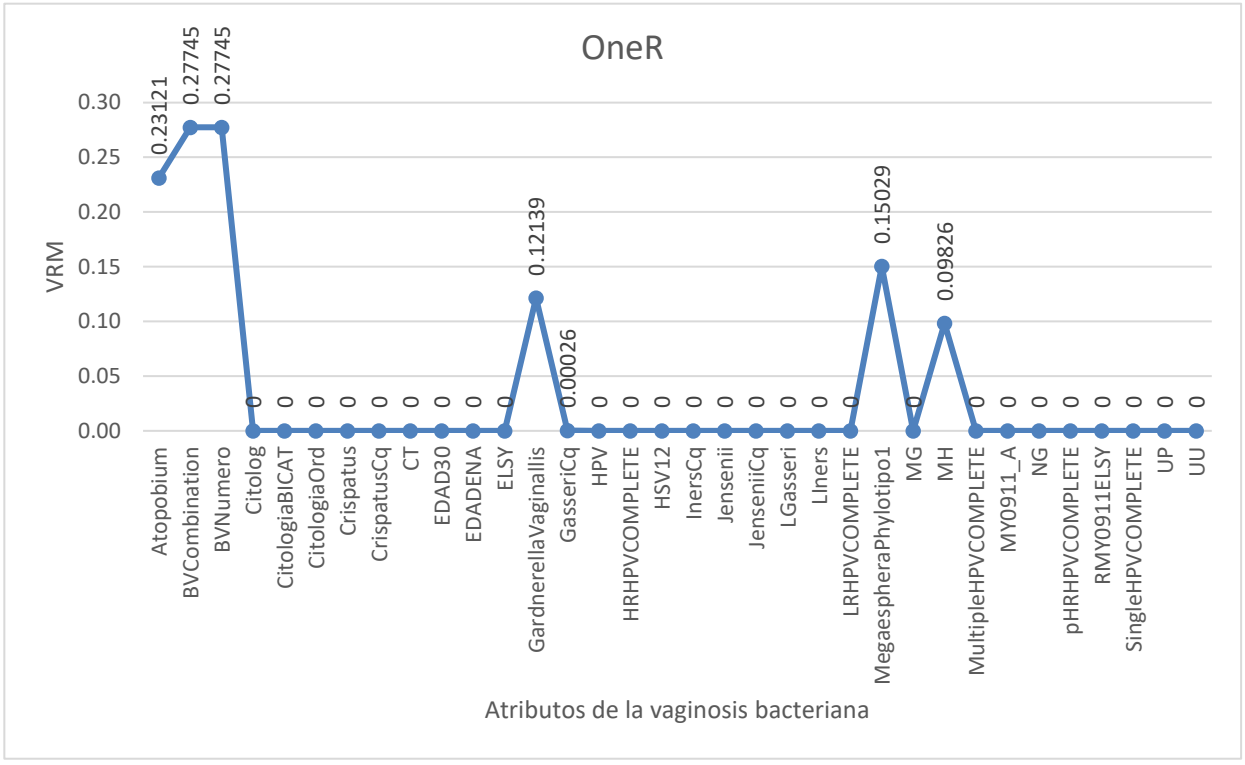

Figura 26. Valor de relevancia media (VRM) obtenida por los atributos de la vaginosis bacteriana en las 30 corridas del método *OneR*.

A partir de estos resultados, y al ordenar los atributos en forma descendente de acuerdo al VRM obtenido mediante *OneR*, se crea el denominado “ranking individual de atributos” para este método. Dicho ranking se muestra en la Tabla 23.

Tabla 23. Ranking individual de atributos de vaginosis bacteriana generado por el método *OneR* mediante las 30 corridas bajo esquema de validación cruzada. Los atributos se ordenan de forma descendente de acuerdo al Valor de Relevancia Medio (VRM) -promedio de las 30 corridas-.

#Rank	Atributo	VRM	#Rank	Atributo	VRM
1	BVCombination	0.2775	18	HSV12	0.0000
2	BVNumero	0.2775	19	InersCq	0.0000
3	Atopobium	0.2312	20	Jensenii	0.0000
4	MegaespheraPhylotipo1	0.1503	21	JenseniiCq	0.0000
5	GardnerellaVaginallis	0.1214	22	LGasseri	0.0000
6	MH	0.0983	23	LIners	0.0000
7	GasseriCq	0.0003	24	LRHPVCOMPLETE	0.0000
8	Crispatus	0.0000	25	MG	0.0000
9	Citolog	0.0000	26	MultipleHPVCOMPLETE	0.0000
10	CitologiaBICAT	0.0000	27	MY0911_A	0.0000
11	CitologiaOrd	0.0000	28	NG	0.0000
12	CrispatusCq	0.0000	29	pHRHPVCOMPLETE	0.0000
13	CT	0.0000	30	RMY0911ELSY	0.0000
14	EDAD30	0.0000	31	SingleHPVCOMPLETE	0.0000
15	ELSY	0.0000	32	UP	0.0000
16	HPV	0.0000	33	UU	0.0000
17	HRHPVCOMPLETE	0.0000	34	EDADENA	0.0000

Finalmente, *OneR* identificó como más relevantes 5 de 34 atributos de la vaginosis bacteriana. Entre estos se destacan atributos como “BVCombination”, “BVNumero”, “Atopobium”, “MegaespheraPhylotipo1”, “GardnerellaVaginallis”, “MH” y “GasseriCq”.

4.2.20 LASSO

La identificación de los atributos más relevantes de la vaginosis bacteriana mediante el algoritmo *LASSO* se realizó de la siguiente manera. Se realizaron 30 corridas del método bajo un esquema de validación cruzada de 10 pliegues. En cada iteración del esquema de validación cruzada se utilizó el 90% de las instancias para crear un conjunto de entrenamiento, la cual es aleatoriamente cambiado en cada iteración. Por cada una de las iteraciones, se calculó en nivel de relevancia de cada atributo del conjunto de entrenamiento con base en el método *LASSO* tal como se describió anteriormente. En cada corrida, se obtuvo un valor de relevancia para cada atributo al promediar las 10 medidas obtenidas en las iteraciones. Al final, las 30 medidas de relevancia se promediaron, con lo cual se obtuvo un valor de relevancia medio (VRM). Cada corrida fue implementada con semillas diferentes para asegurar la aleatoriedad de los datos. En la Figura 27 se muestran los VRM resultantes de cada atributo.

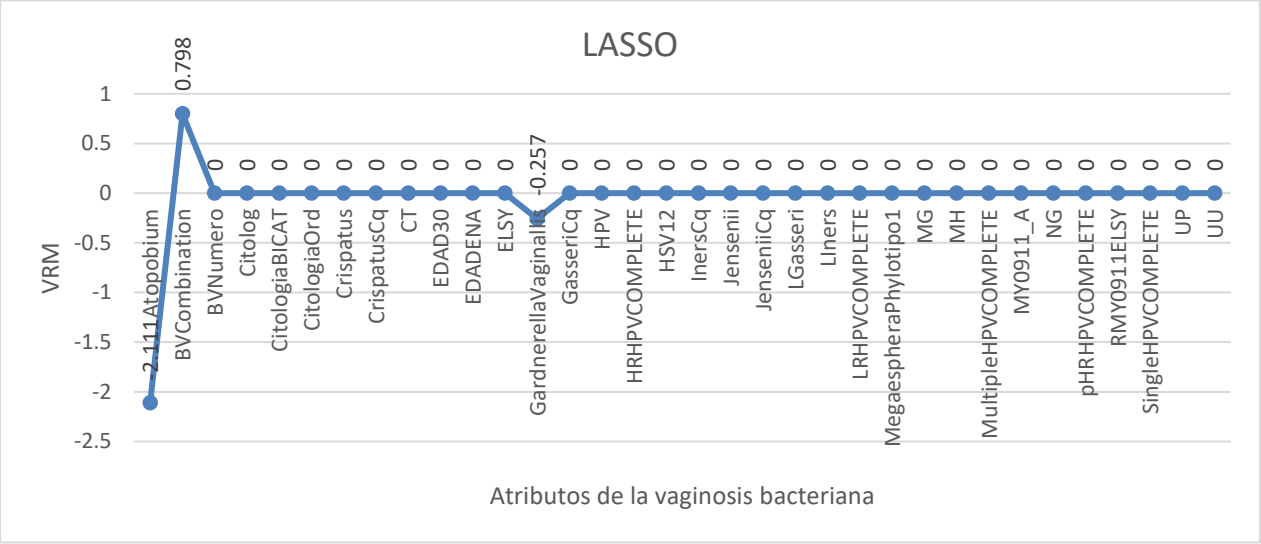

Figura 27. Valor de relevancia media (VRM) obtenida por los atributos de la vaginosis bacteriana en las 30 corridas del método *LASSO.*

Con base en los resultados obtenidos, se crea el ranking individual de atributos mediante el método *LASSO.* Para ello, los atributos en el conjunto de datos son ordenados de forma descendente al considerar el VRM obtenido de las 30 corridas como criterio de ordenamiento. El ranking individual de atributos de la vaginosis bacteriana basado en el método *LASSO* se muestra en la Tabla 24.

Tabla 24. Ranking individual de atributos de vaginosis bacteriana generado mediante el método LASSO. Los atributos se ordenan de forma descendente de acuerdo al Valor de Relevancia Medio (VRM) -promedio de los 30 resultados-.

#Rank	Atributos	VRM	#Rank	Atributos	VRM
1	BVCombination	0.798	18	JenseniiCq	0
2	BVNumero	0	19	LGasseri	0
3	Citolog	0	20	LIners	0
4	CitologiaBICAT	0	21	LRHPVCOMPLETE	0
5	CitologiaOrd	0	22	MegaespheraPhylotipo1	0
6	Crispatus	0	23	MG	0
7	CrispatusCq	0	24	MH	0
8	CT	0	25	MultipleHPVCOMPLETE	0
9	EDAD30	0	26	MY0911_A	0
10	EDADENA	0	27	NG	0
11	ELSY	0	28	pHRHPVCOMPLETE	0
12	GasseriCq	0	29	RMY0911ELSY	0
13	HPV	0	30	SingleHPVCOMPLETE	0
14	HRHPVCOMPLETE	0	31	UP	0
15	HSV12	0	32	UU	0
16	InersCq	0	33	GardnerellaVaginallis	-0.257
17	Jensenii	0	34	Atopobium	-2.111

En este caso, el método *LASSO* identificó a “BVCombination” como el atributo más relevante de la VB. Sin embargo, este método también consideró a “GardnerellaVaginallis” y “Atopobium” como atributos relevantes, pero con coeficientes en valores negativos.

4.2.21 Bosques aleatorios regularizado (RRF)

El ranking con los atributos más relevantes para el diagnóstico de la vaginosis bacteriana mediante el algoritmo *RRF* fue calculado de la siguiente manera. Se

realizaron 30 corridas del método bajo un esquema de validación cruzada de 10 pliegues. En cada iteración del esquema de validación cruzada se utilizó el 90% de las instancias para crear un conjunto de entrenamiento, sobre el cual se aplicó el método. Las instancias de entrenamiento son aleatoriamente cambiadas en cada iteración. Por cada una de las iteraciones, se calculó en nivel de relevancia de cada atributo del conjunto de entrenamiento basado en el criterio de ranqueo del algoritmo *RRF* descrito anteriormente. En cada corrida, se obtuvo un valor de relevancia para cada atributo al promediar las 10 medidas obtenidas en las iteraciones. Después, las 30 medidas de relevancia fueron promediadas, con lo cual se obtuvo un valor de relevancia medio (VRM). Este valor es utilizado como criterio de ranqueo de los atributos. Cada corrida fue implementada con semillas diferentes para asegurar la aleatoriedad de los datos. En la Figura 28 se muestran los VRM resultantes de cada atributo obtenido mediante el método *RRF.*

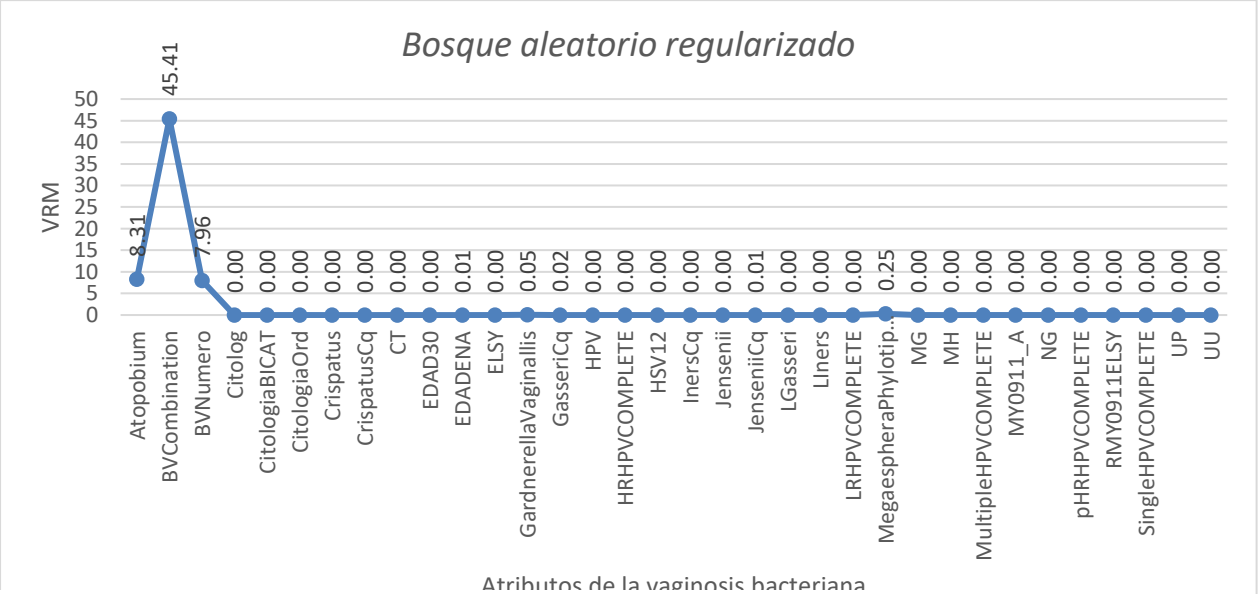

Figura 28. Valor de relevancia media (VRM) obtenida por los atributos de la vaginosis bacteriana en las 30 corridas del método *bosques aleatorios regularizado.*

Con base en estos resultados, se crea el ranking individual de atributos mediante el método *RRF*. Para ello, los atributos en el conjunto de datos son ordenados de forma descendente respecto al VRM obtenido de las 30 corridas. El ranking individual de atributos de la vaginosis bacteriana basado en el método *Regularized Random Forest* se muestra en la Tabla 25.

Tabla 25. Ranking individual de atributos de vaginosis bacteriana generado mediante el método *Regularized Random Forest (RRF)*. Los atributos se ordenan de forma descendente de acuerdo al Valor de Relevancia Medio (VRM) -promedio de los 30 resultados-.

#Rank	Features	VRM	#Rank	Features	VRM
1	BVCombination	45.4063691	18	CitologiaOrd	0.0000128
2	Atopobium	8.31473725	19	CT	0
3	BVNumero	7.95801566	20	ELSY	0
4	MegaespheraPhylotipo1	0.24950915	21	HPV	0
5	GardnerellaVaginallis	0.04898354	22	HRHPVCOMPLETE	0
6	GasseriCq	0.02115138	23	Jensenii	0
7	EDADENA	0.00715599	24	LGasseri	0
8	JenseniiCq	0.00644756	25	LIners	0
9	CrispatusCq	0.00235881	26	LRHPVCOMPLETE	0
10	HSV12	0.00209434	27	MG	0
11	MH	0.00176112	28	MultipleHPVCOMPLETE	0
12	Citolog	0.00064763	29	MY0911_A	0
13	InersCq	0.00032282	30	NG	0
14	Crispatus	0.00024742	31	pHRHPVCOMPLETE	0
15	EDAD30	0.0001898	32	RMY0911ELSY	0
16	CitologiaBICAT	0.00018545	33	SingleHPVCOMPLETE	0
17	UP	0.00013333	34	UU	0

En este caso, cabe destacar aquellos atributos que obtuvieron un VRM por encima de la media aritmética (1.82412), estos son solamente tres: “BVCombination”, “Atopobium” y “BVNumero”.

4.3 Primer ranking general (Media aritmética)

Con base en los resultados obtenidos de los rankings de atributos individuales en la Sección 4.2 se creó un primer ranking general de atributos. Para esto, fueron considerados aquellos rankings individuales creados a partir de los métodos selectores de atributos que determinan el nivel de relevancia de los atributos mediante un valor

numérico. Primero, se reescalaron los valores VRM obtenidos por los atributos en los rankings individuales para estandarizar los valores entre el 0 y el 1. Después, todos los VRM reescalados se promediaron para obtener el valor medio al que se le denominó Valor de Relevancia General (VRG). Este valor representa la medida de relevancia de cada atributo en el ranking general de atributos. Para ilustrar de mejor forma la manera en que el primer ranking general de atributos fue obtenido, se presenta la Tabla 26.

Tabla 26. Estructura en formato de tabla para el cálculo del ranking general de atributos

Atributos	**Método 1**	**Método 2**	**Método 3**	**…**	**Método *N***	**VRG**
Atributo 1	VRM	VRM	VRM	…	VRM	VRG del Atributo 1
Atributo 2	VRM	VRM	VRM	…	VRM	VRG del Atributo 2
Atributo 3	VRM	VRM	VRM	…	VRM	VRG del Atributo 3
Atributo 4	VRM	VRM	VRM	…	VRM	VRG del Atributo 4
…	…	…	…	…	…	…
Atributo *N*	VRM	VRM	VRM	…	VRM	VRG del Atributo *N*

Al implementar la estructura anterior y calcular el valor de relevancia medio de los atributos de la VB calculados en los rankings individuales de atributos en la sección 4.2, se obtuvo un valor de relevancia general (VRG) para cada atributo. En la Tabla 27 se muestran los valores obtenidos y calculados en este paso.

Tabla 27. Cálculo del primer ranking general de atributos de la vaginosis bacteriana a partir de los rankings individuales de atributos.

MSA[2] / Atributos	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	VRG
Atopobium	0.842	0.002	0.887	0.897	0.156	0.083	0.600	0.571	0.084	0.887	0.887	0.698	0.683	1.000	0.897	0.434	0.601
BVCombination	0.382	0.003	1.000	1.000	0.239	0.454	0.852	0.795	0.034	0.847	0.984	1.000	-1.057	0.017	1.000	1.000	0.535
BVNumero	0.497	0.003	1.000	0.696	0.138	0.080	0.852	0.481	0.013	0.537	0.620	0.794	0.189	0.609	0.696	0.272	0.467
Citolog	0.044	0.000	0.000	0.111	0.006	0.000	0.000	-0.006	0.000	0.088	0.095	0.000	-0.014	0.003	0.111	0.008	0.028
CitologiaBICAT	0.070	0.000	0.000	0.127	0.010	0.000	0.000	-0.006	0.000	0.114	0.114	0.000	0.048	0.004	0.127	0.005	0.038
CitologiaOrd	0.044	0.000	0.000	0.111	0.004	0.000	0.000	-0.006	0.000	0.088	0.095	0.000	0.014	0.000	0.111	0.007	0.029
Crispatus	0.052	0.000	0.009	0.240	0.029	0.000	0.048	0.030	0.002	0.240	0.240	0.000	0.035	0.003	0.240	0.009	0.073
CrispatusCq	-0.004	0.000	0.000	0.121	0.028	0.000	0.000	0.015	0.000	0.077	0.094	0.000	0.002	0.001	0.121	0.025	0.030
CT	0.011	0.000	0.000	0.075	0.023	0.000	0.021	0.018	0.001	0.201	0.201	0.039	0.050	0.003	0.075	0.009	0.045
EDAD30	0.033	0.000	0.000	0.058	-0.003	0.000	0.000	-0.013	0.000	0.053	0.053	0.000	-0.059	0.005	0.059	0.005	0.012
EDADENA	0.000	0.000	0.006	0.176	0.021	0.000	0.014	0.028	0.001	0.156	0.136	0.007	0.006	0.003	0.176	0.023	0.047
ELSY	0.021	0.000	0.000	0.112	0.002	0.000	0.000	-0.004	0.000	0.119	0.119	0.000	0.034	0.005	0.113	0.006	0.033
GardnerellaVaginallis	0.459	0.001	0.600	0.579	0.107	0.000	0.244	0.226	0.011	0.599	0.599	0.295	0.662	0.721	0.579	0.082	0.360
GasseriCq	0.005	0.000	0.002	0.053	0.010	0.000	0.000	0.028	0.000	0.023	0.044	0.000	0.000	0.008	0.053	0.023	0.016
HPV	-0.001	0.000	0.000	0.093	0.000	0.000	0.000	-0.010	0.000	0.087	0.087	0.000	-0.034	0.008	0.094	0.004	0.021
HRHPVCOMPLETE	0.009	0.000	0.000	0.066	0.004	0.000	0.000	-0.012	0.000	0.065	0.065	0.000	0.034	0.004	0.066	0.004	0.019
HSV12	0.015	0.000	0.000	0.026	-0.003	0.000	0.000	-0.001	0.000	0.077	0.077	0.001	-0.006	0.013	0.026	0.003	0.014
InersCq	0.008	0.000	0.000	0.168	0.011	0.000	0.000	0.025	0.000	0.125	0.130	0.000	-0.001	0.012	0.168	0.025	0.042
Jensenii	0.010	0.000	0.000	0.122	0.018	0.000	0.007	0.004	0.001	0.152	0.152	0.008	0.053	0.012	0.121	0.005	0.041
JenseniiCq	0.032	0.000	0.000	0.054	-0.001	0.000	0.000	0.003	0.000	0.039	0.042	0.000	-0.004	0.005	0.052	0.018	0.015

[2] A:Relief; B: OneR; C: Chi.Square; D: J48; E: Boruta; F: RRF; G: InfGain; H: MDL; I: Fisher Score; J:Pearson; K: Spearman; L: Incertidumbre simétrica Uncertainty; M: SVM; N: RL; O: DT relevancia; P: RF.

Tabla 27. (Continuación)

MSA[2] Atributos	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	VRG
LGasseri	0.010	0.000	0.000	0.075	0.014	0.000	0.003	0.000	0.001	0.127	0.127	0.006	0.069	0.009	0.075	0.004	0.033
Lners	0.016	0.000	0.000	0.034	-0.004	0.000	0.000	-0.015	0.000	0.028	0.028	0.000	-0.011	0.004	0.032	0.004	0.007
LRHPVCOMPLETE	0.037	0.000	0.000	0.040	-0.002	0.000	0.000	-0.010	0.000	0.051	0.051	0.000	-0.041	0.004	0.040	0.004	0.011
MegaespheraPhylotipo1	0.398	0.002	0.669	0.580	0.118	0.002	0.305	0.285	0.013	0.669	0.669	0.395	0.001	0.509	0.580	0.117	0.332
MG	-0.004	0.000	0.000	0.015	-0.004	0.000	0.000	-0.002	0.000	0.063	0.063	0.000	0.000	0.000	0.014	0.002	0.009
MH	0.178	0.001	0.521	0.380	0.069	0.000	0.181	0.163	0.006	0.521	0.521	0.056	0.058	0.002	0.380	0.049	0.193
MultipleHPVCOMPLETE	0.011	0.000	0.000	0.037	0.001	0.000	0.000	-0.012	0.000	0.043	0.043	0.000	0.021	0.011	0.039	0.004	0.012
MY0911_A	0.040	0.000	0.000	0.022	0.004	0.000	0.000	-0.015	0.000	0.022	0.022	0.000	-0.024	0.003	0.025	0.005	0.007
NG	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.003	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000
pHRHPVCOMPLETE	-0.013	0.000	0.000	0.017	-0.004	0.000	0.000	-0.008	0.000	0.037	0.037	0.000	0.007	0.007	0.018	0.002	0.006
RMY0911ELSY	0.016	0.000	0.000	0.024	0.002	0.000	0.000	-0.004	0.000	0.020	0.020	0.000	-0.024	0.009	0.024	0.006	0.006
SingleHPVCOMPLETE	0.002	0.000	0.000	0.064	-0.003	0.000	0.000	-0.008	0.000	0.078	0.078	0.000	0.014	0.010	0.064	0.004	0.019
UP	0.031	0.000	0.000	0.158	0.000	0.000	0.001	-0.001	0.001	0.143	0.143	0.000	-0.018	0.002	0.158	0.007	0.039
UU	0.001	0.000	0.000	0.035	-0.003	0.000	0.000	-0.006	0.000	0.069	0.069	0.000	0.048	0.000	0.036	0.002	0.016

En la Figura 29 se muestran gráficamente los niveles de relevancia general obtenidos por cada atributo de la vaginosis bacteriana de acuerdo con los niveles de relevancia medio obtenidos de los rankings individuales de atributos resultantes de la Sección 4.2.

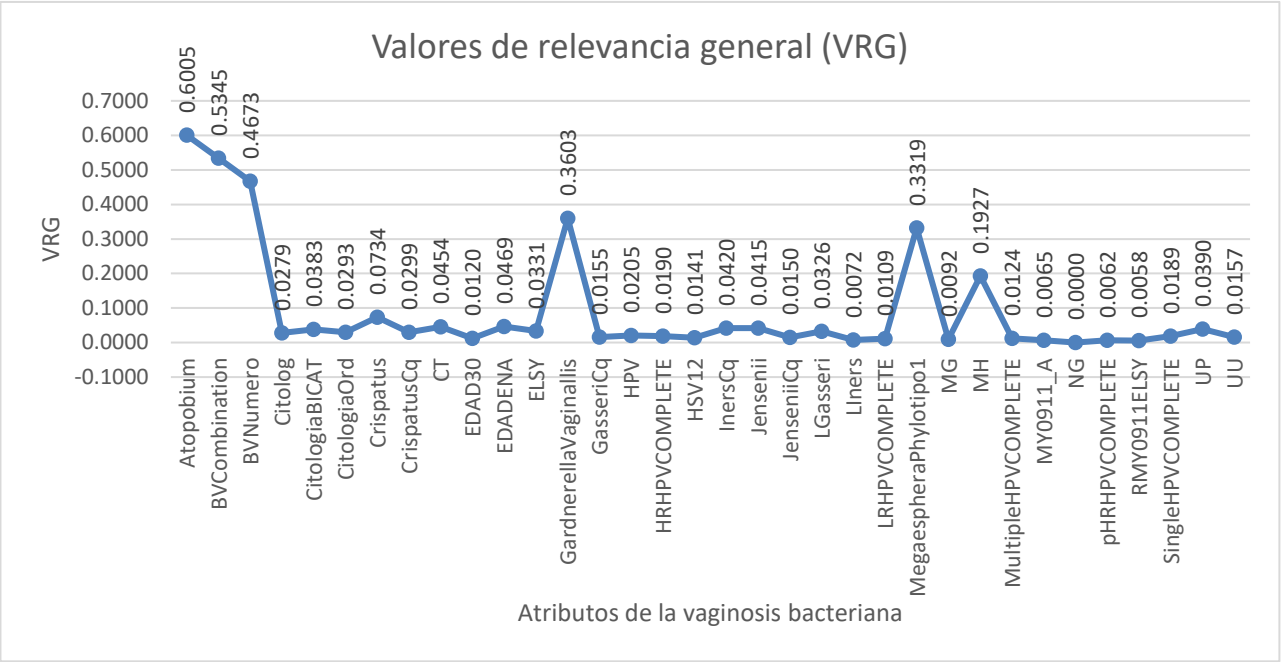

Figura 29. Valores de relevancia general (VRG) obtenido por los atributos de la vaginosis bacteriana con base en los valores de relevancia medio (VRM) mediante el cálculo de los rankings individuales de atributos.

Con base en estos resultados, se crea el primer ranking general de atributos mostrado en la Tabla 28. Para ello, los atributos se ordenaron en forma descendente considerando el VRG obtenido con anterioridad.

Tabla 28. Primer ranking general de atributos de la vaginosis bacteriana con base en el valor de relevancia general (VRG) calculado a partir de los métodos de selección de atributos.

#Rank	Atributo	VRM	#Rank	Atributo	VRM
1	Atopobium	0.6005	18	Citolog	0.0279
2	BVCombination	0.5345	19	HPV	0.0205
3	BVNumero	0.4673	20	HRHPVCOMPLETE	0.0190
4	GardnerellaVaginallis	0.3603	21	SingleHPVCOMPLETE	0.0189
5	MegaespheraPhylotipo1	0.3319	22	UU	0.0157
6	MH	0.1927	23	GasseriCq	0.0155
7	Crispatus	0.0734	24	JenseniiCq	0.0150
8	EDADENA	0.0469	25	HSV12	0.0141
9	CT	0.0454	26	MultipleHPVCOMPLETE	0.0124
10	InersCq	0.0420	27	EDAD30	0.0120
11	Jensenii	0.0415	28	LRHPVCOMPLETE	0.0109
12	UP	0.0390	29	MG	0.0092
13	CitologiaBICAT	0.0383	30	LIners	0.0072
14	ELSY	0.0331	31	MY0911_A	0.0065
15	LGasseri	0.0326	32	pHRHPVCOMPLETE	0.0062
16	CrispatusCq	0.0299	33	RMY0911ELSY	0.0058
17	CitologiaOrd	0.0293	34	NG	0.0000

El primer ranking general de atributos muestra el nivel de relevancia promedio de los atributos para el diagnóstico de la vaginosis bacteriana con base en los métodos de selección de atributos implementados. Cada método selector de atributos considera el nivel de relevancia de acuerdo a diferentes criterios de evaluación detallados en Sección 2.4. El escalado y promedio de dichos valores de relevancia, permitieron crear el ranking mostrado con anterioridad. El primer ranking general muestra atributos de la vaginosis bacteriana tales como “Atopobium”, “BVCombination”, “BVNumero”, “GardnerellaVaginallis”, “MegaespheraPhylotipo1”, “MH” con altos niveles de relevancia (VRG) por encima de la media (0.0927). Estos atributos destacan entre los demás como los más relacionados al diagnóstico de la vaginosis bacteriana.

4.4 Segundo ranking general (Análisis de frecuencias)

Un segundo ranking general de atributos de la vaginosis bacteriana se calculó a partir de una distribución de frecuencias. Un análisis, tabla o distribución de frecuencia estadística se basa en la idea de que ciertos eventos o un conjunto de ellos aparecen más a menudo que

otras. En este ranking se consideraron todos los rankings individuales resultantes de la Sección 4.2. Para obtenerlo, se calcularon las posiciones obtenidas por los atributos en cada uno de los rankings individuales de atributos y con base en ello se calculó la moda estadística de las posiciones obtenidas por los atributos. Para ilustrar de mejor forma la manera en que la distribución de frecuencias fue obtenida, se presenta la Tabla 29.

Tabla 29. Estructura para la obtención del análisis de frecuencias de atributos.

Atributos	**Método 1**	**Método 2**	**Método 3**	**…**	**Método *N***	**Moda**
Atributo 1	Posición	Posición	Posición	…	Posición	Moda del Atributo 1
Atributo 2	Posición	Posición	Posición	…	Posición	Moda del Atributo 2
Atributo 3	Posición	Posición	Posición	…	Posición	Moda del Atributo 3
Atributo 4	Posición	Posición	Posición	…	Posición	Moda del Atributo 4
…	…	…	…	…	…	…
Atributo *N*	Posición	Posición	Posición	…	Posición	Moda del Atributo *N*

Con la utilización de la estructura anterior, se calculó la Tabla 30. En dicha tabla se muestran las posiciones de los atributos de la VB obtenidos en los rankings individuales en la sección 4.2. Al final, se calculó la moda estadística de las posiciones (última columna) para definir la posición en los rankings de atributos individuales con mayor frecuencia.

Tabla 30. Distribución de frecuencias de las posiciones de los atributos obtenidas en los rankings individuales. MSA: métodos selectores de atributos.

Atributos \ MSA[3]	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	Moda
Atopobium	1	3	3	2	2	2	2	2	1	2	2	3	1	1	1	N/A	N/A	3	4	N/A	N/A	2
BVCombination	5	1	1	1	1	1	1	1	2	1	1	1	34	5	2	1	N/A	1	2	1	N/A	1
BVNumero	2	2	2	3	3	5	3	3	3	15	4	2	3	3	N/A	N/A	1	2	3		1	3
Citolog	9	9	10	16	16	13	16	23	16	18	16	N/A	27	25	N/A	N/A	N/A	N/A	N/A	N/A	N/A	16
CitologiaBICAT	7	10	11	11	15	21	15	24	15	23	15	N/A	8	22	N/A	N/A	N/A	N/A	N/A	N/A	N/A	15
CitologiaOrd	10	11	12	15	19	20	19	25	17	29	17	N/A	14	32	N/A	N/A	N/A	N/A	N/A	N/A	N/A	19
Crispatus	8	8	7	7	7	15	7	7	7	7	7	N/A	10	23	N/A	N/A	N/A	N/A	N/A	N/A	N/A	7
CrispatusCq	33	12	13	13	8	9	8	12	20	12	18	N/A	19	30	N/A	N/A	N/A	N/A	N/A	N/A	N/A	12
CT	20	13	14	18	9	25	9	11	10	8	8	7	7	27	N/A	N/A	N/A	N/A	N/A	N/A	N/A	13
EDAD30	13	14	15	22	30	22	30	32	24	33	25	N/A	33	17	N/A	N/A	N/A	N/A	N/A	N/A	N/A	22
EDADENA	30	34	8	8	10	10	10	9	9	9	11	9	17	26	N/A	N/A	N/A	N/A	1	1	N/A	10
ELSY	16	15	16	14	21	16	21	20	14	14	14	N/A	11	16	N/A	N/A	N/A	N/A	N/A	N/A	N/A	14
GardnerellaVaginallis	3	5	5	5	5	4	5	5	5	16	5	5	2	2	3	N/A	N/A	N/A	N/A	N/A	N/A	5
GasseriCq	26	7	9	24	14	6	14	8	32	20	27	N/A	21	14	N/A	N/A	N/A	N/A	N/A	N/A	N/A	14
HPV	31	16	17	17	23	N/A	23	28	18	13	19	N/A	31	13	N/A	N/A	N/A	N/A	N/A	N/A	N/A	17
HRHPVCOMPLETE	24	17	18	20	18	14	18	30	22	27	23	N/A	12	19	N/A	N/A	N/A	N/A	N/A	N/A	N/A	18
HSV12	19	18	19	29	29	N/A	29	17	21	26	21	11	25	6	N/A	N/A	N/A	N/A	N/A	N/A	N/A	29
InersCq	25	19	20	9	13	7	13	10	13	5	12	N/A	23	7	N/A	N/A	N/A	N/A	N/A	N/A	N/A	13
Jensenii	23	20	21	12	11	19	11	13	8	30	9	8	6	8	N/A	N/A	N/A	N/A	N/A	N/A	N/A	8
JenseniiCq	14	21	22	23	26	8	26	14	28	17	29	N/A	24	18	N/A	N/A	N/A	N/A	N/A	N/A	N/A	14
LGasseri	22	22	23	19	12	N/A	12	15	11	21	13	10	4	11	N/A	N/A	N/A	N/A	N/A	N/A	N/A	22
LIners	18	23	24	28	34	17	34	33	30	32	31	N/A	26	20	N/A	N/A	N/A	N/A	N/A	N/A	N/A	23
LRHPVCOMPLETE	12	24	25	25	27	N/A	27	29	26	6	26	N/A	32	21	N/A	N/A	N/A	N/A	N/A	N/A	N/A	25
MegaespheraPhylotipo1	4	4	4	4	4	3	4	4	4	3	3	4	20	4	N/A	N/A	N/A	N/A	N/A	N/A	N/A	4
MG	32	25	26	33	33	N/A	33	18	25	4	24	N/A	22	33	N/A	N/A	N/A	N/A	N/A	N/A	N/A	33
MH	6	6	6	6	6	11	6	6	6	19	6	6	5	28	N/A	N/A	N/A	N/A	N/A	N/A	N/A	6
MultipleHPVCOMPLETE	21	26	27	26	22	N/A	22	31	27	22	28	N/A	13	9	N/A	N/A	N/A	N/A	N/A	N/A	N/A	26
MY0911_A	11	27	28	31	17	23	17	34	31	31	32	N/A	29	24	N/A	N/A	N/A	N/A	N/A	N/A	N/A	31
NG	29	28	29	34	25	N/A	25	19	34	28	34	N/A	18	34	N/A	N/A	N/A	N/A	N/A	N/A	N/A	34
pHRHPVCOMPLETE	34	29	30	32	32	N/A	32	26	29	25	30	N/A	16	15	N/A	N/A	N/A	N/A	N/A	N/A	N/A	32

[3] A:Relief; B:OneR; C:Chi cuadrada; D:Entropía; E:Boruta; F:RRF; G:Ganancia de información; H:MDL; I:Puntuación de Fisher; J:Pearson; K:Spearman; L:Incertidumbre simétrica; M: SVM; N:RL; O:LASSO; P:CFS; Q:SFS; R:SBS; S:SFFS; T:SBFS; U:Consistencia.

Tabla 30. (Continuación)

Atributo	MSA[3]	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	Moda
RMY0911ELSY		17	30	31	30	20	12	20	21	33	24	33	N/A	30	12	N/A	N/A	N/A	N/A	N/A	N/A	N/A	30
SingleHPVCOMPLETE		27	31	32	21	31	18	31	27	19	11	20	N/A	15	10	N/A	N/A	N/A	N/A	N/A	N/A	N/A	27
UP		15	32	33	10	24	24	24	16	12	10	10	N/A	28	29	N/A	N/A	N/A	N/A	N/A	N/A	N/A	10
UU		28	33	34	27	28	N/A	28	22	23	34	22	N/A	9	31	N/A	N/A	N/A	N/A	N/A	N/A	N/A	28

En la Figura 30 se muestra gráficamente los resultados de la tabla de frecuencias mostrada con anterioridad.

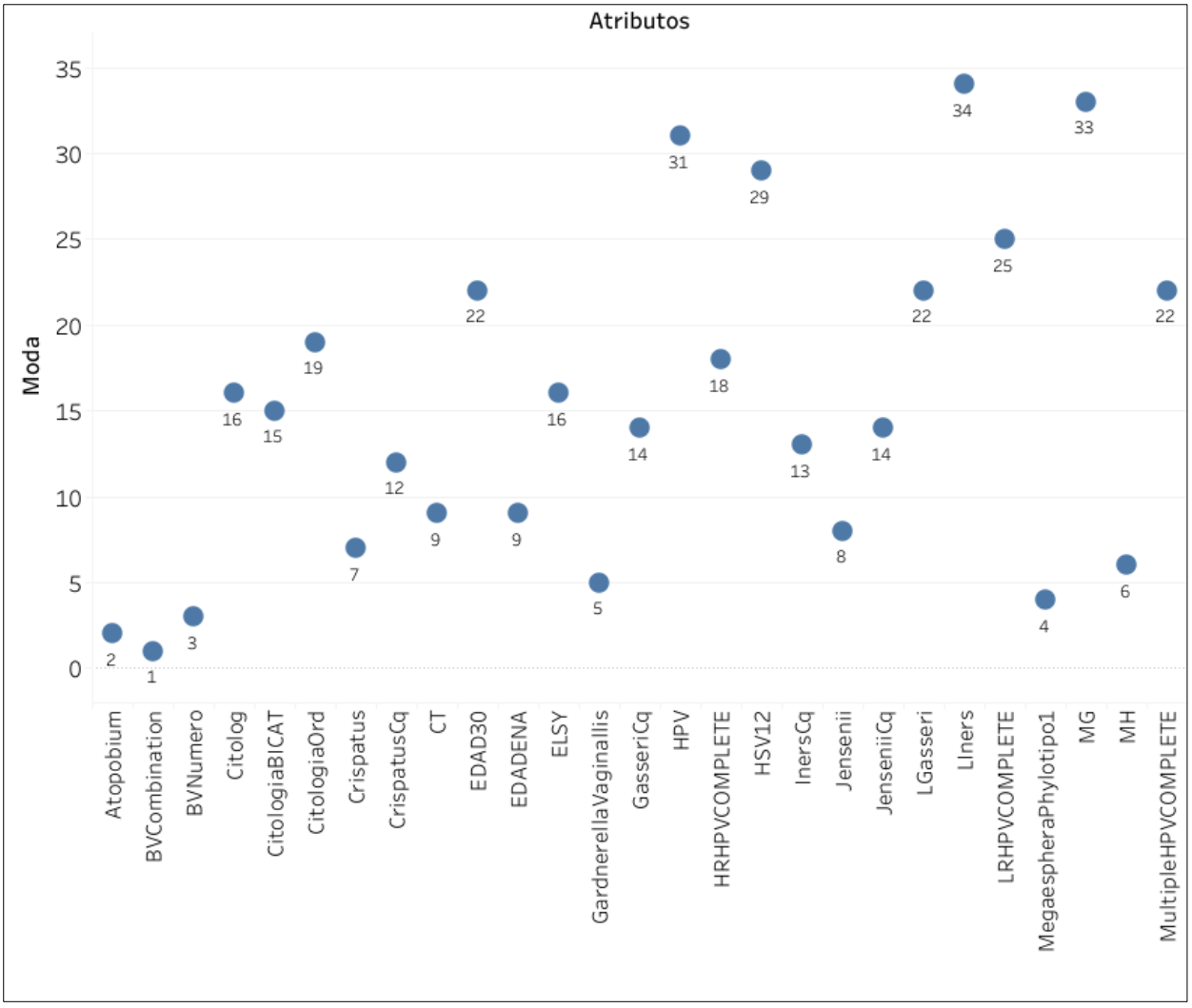

Figura 30. Moda estadística calculada para cada atributo de la vaginosis bacteriana (VB) con base en un análisis de distribución de frecuencias obtenida de los rankings individuales de atributos

Con base en los resultados anteriores, se crea el segundo ranking de atributos. Para ello, se ordenaron los atributos en forma descendente de acuerdo a la moda estadística. En la Tabla 31 se muestra el segundo ranking general de atributos de la vaginosis bacteriana.

Tabla 31. Segundo ranking general de atributos de la vaginosis bacteriana calculado con base en un análisis de distribución de frecuencias a partir de los resultados de los rankings individuales de atributos.

#Rank	Atributos	Moda	#Rank	Atributos	Moda
1	BVCombination	1	18	Citolog	16
2	Atopobium	2	19	HPV	17
3	BVNumero	3	20	HRHPVCOMPLETE	18
4	MegaespheraPhylotipo1	4	21	CitologiaOrd	19
5	GardnerellaVaginallis	5	22	EDAD30	22
6	MH	6	23	LGasseri	22
7	Crispatus	7	24	LIners	23
8	Jensenii	8	25	LRHPVCOMPLETE	25
9	EDADENA	10	26	MultipleHPVCOMPLETE	26
10	UP	10	27	SingleHPVCOMPLETE	27
11	CrispatusCq	12	28	UU	28
12	CT	13	29	HSV12	29
13	InersCq	13	30	RMY0911ELSY	30
14	ELSY	14	31	MY0911_A	31
15	GasseriCq	14	32	pHRHPVCOMPLETE	32
16	JenseniiCq	14	33	MG	33
17	CitologiaBICAT	15	34	NG	34

En este segundo ranking, el atributo definido como el más relevante para la vaginosis bacteriana es “BVCombination”, seguido por atributos tales como “Atopobium”, “BVNumero”, “MegaespheraPhylotipo1”, “GardnerellaVaginallis” y “MH” que ocupan del segundo al sexto lugar respectivamente en el ranking.

Capítulo 5. Modelos predictivos de la vaginosis bacteriana

Los algoritmos de clasificación consisten de una fase de aprendizaje, donde un modelo de clasificación se construye y entrena, y una fase de clasificación, donde el modelo se utiliza para predecir una clase respecto a los datos proporcionados [42]. De acuerdo con Aggarwal [1], el problema de clasificación se puede explicar de la siguiente manera: dado un conjunto de puntos de datos de entrenamiento, junto con etiquetas de datos de entrenamiento asociadas, el problema es determinar la etiqueta para una instancia de prueba no etiquetada. En este trabajo se implementaron tres algoritmos de clasificación para determinar la capacidad de los modelos de predecir entre pacientes con VB positiva y pacientes con VB negativa: máquinas de vectores de soporte, regresión logística y arboles de decisión. La manera en que trabaja cada uno de estos métodos clasificadores basados en aprendizaje maquina se detalla en la sección 2.3.

Los experimentos con modelos de clasificación para evaluar la capacidad de predicción de la vaginosis bacteriana se muestran gráficamente en la Figura 31.

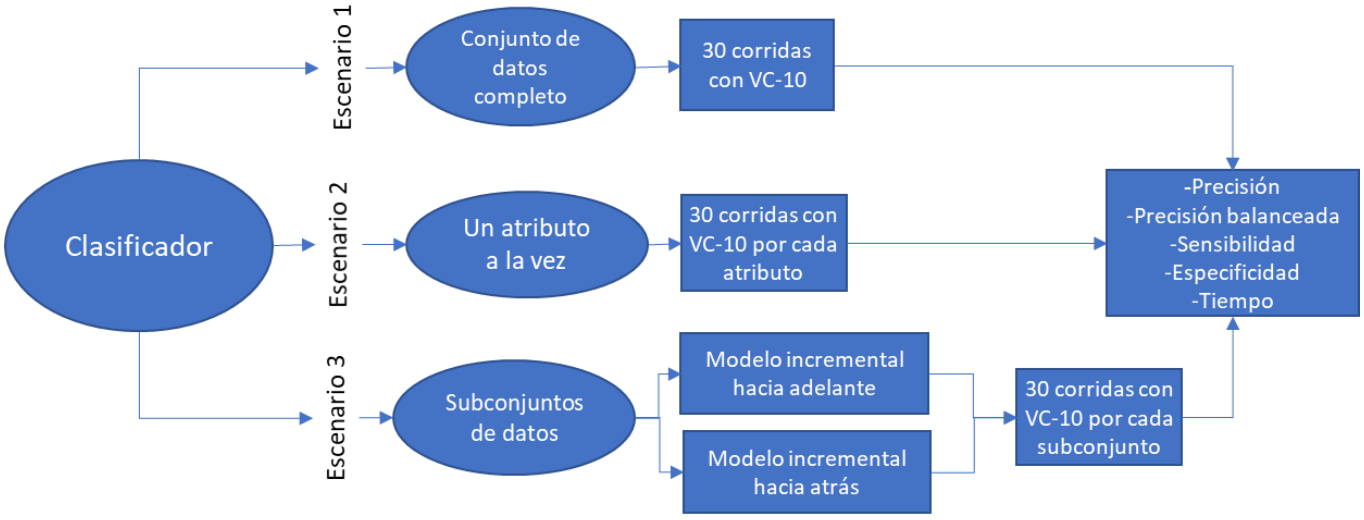

Figura 31. Modelo experimental para el diagnóstico de la vaginosis bacteriana.

Cada uno de los tres escenarios implementados se detallan en las siguientes secciones.

5.1 Escenario experimental uno

El primer escenario consistió de 30 corridas por cada método clasificador utilizando el conjunto de datos completo de la vaginosis bacteriana descrito en la sección 4.1. Cada corrida se condujo bajo un esquema de VC-10, el cual se describe en la sección 2.5. A través de las 30 corridas se utilizó una semilla diferente para asegurar la división aleatoria de los datos. Particularmente en este escenario se utilizó el conjunto de atributos completo, es decir, 34 atributos y la etiqueta de clase. Se obtuvo un promedio de cada métrica de rendimiento descritos en la sección 2.5 con base en las 30 corridas con VC-10. Este promedio se considera como el rendimiento final de los modelos predictivos en este escenario.

Con base en los experimentos realizados, en la Tabla 32 se proporcionan los resultados obtenidos por los modelos de clasificación de la vaginosis bacteriana en el escenario uno.

Tabla 32. Rendimiento predictivo general obtenido por los tres métodos de clasificación experimentados. SVM: support vector machine, RL: regresión logística, DT: decision tree, RF: random forests; Ms: microsegundos

Clasificador	Atributos	Precisión	Precisión balanceada	Sensibilidad	Especificidad	Tiempo (Ms)
SVM	34	1	1	1	1	1.1447
RL	34	0.9998	0.9996	1	0.9993	1.3447
DT	34	1	1	1	1	11.3197
RF	34	1	1	1	1	5.7052

En este escenario se experimentó con la creación de modelos de clasificación usando el conjunto de atributos completo, es decir, los 34 atributos que integran los datos de la BV. Tanto el SVM, DT y RF obtuvieron los valores más altos posibles en todas las métricas de desempeño. En contraste, RL obtuvo el desempeño más bajo de los tres métodos experimentados. A pesar de ello, su precisión fue cercana al 100%.

Al considerar el tiempo como métrica de desempeño, el mejor rendimiento lo obtuvo el SVM con un promedio de 1.1447 microsegundos, seguido por RL con 1.3447 microsegundos, RF

con 5.7052 microsegundos y DT con 11.3197 microsegundos. Por tanto, el SVM como método clasificador obtuvo el rendimiento más alto como modelo predictivo de la vaginosis bacteriana en el escenario uno.

5.2 Escenario experimental dos

Se realizaron 30 corridas por cada método de clasificación bajo un esquema de VC-10, tal cual como en el escenario uno. A diferencia del primer escenario, en este caso se utilizaron los atributos del conjunto de datos de vaginosis bacteriana uno a la vez. Es decir, un atributo y la clase principal se utilizaron para crear modelos predictivos de la BV. Esto, con el fin de evaluar el rendimiento de los clasificadores para distinguir entre ambas clases de BV usando solamente un atributo de manera individual. Tal cual como en el escenario uno, se calcularon los promedios de las medidas de rendimiento mencionadas en la sección 2.5 con base en las 30 corridas.

Con base en los experimentos realizados, las Tablas 33, 34, 35 y 36 muestran los valores promedio de rendimiento obtenidos de las 30 corridas con VC-10 de los clasificadores SVM, RL, DT y RF respectivamente.

Tabla 33. Rendimiento promedio de las 30 corridas del método *máquina de vector soporte* (SVM) en el escenario dos al utilizar un atributo del conjunto de datos de la vaginosis bacteriana (VB) a la vez como modelo predictivo. Los resultados se ordenan alfabéticamente respecto a la columna "Atributo". Ms: Microsegundos.

Clasificador	Atributo	Precisión	Precisión balanceada	Sensibilidad	Especificidad	Tiempo (Ms)
SVM	Atopobium	0.9534	0.9480	0.9601	0.936	0.7726
SVM	BVCombination	**1**	**1**	1	1	0.7778
SVM	BVNumero	0.8606	0.8709	0.8479	0.894	0.7834
SVM	Citolog	0.7227	0.5	1	0	0.8622
SVM	CitologiaBICAT	0.7227	0.5	1	0	0.864
SVM	CitologiaOrd	0.7227	0.5	1	0	0.8682
SVM	Crispatus	0.7227	0.5	1	0	0.79
SVM	CrispatusCq	0.7227	0.5	1	0	0.8124
SVM	CT	0.7400	0.5389	0.9919	0.086	0.7982
SVM	EDAD30	0.7227	0.5	1	0	0.8022
SVM	EDADENA	0.7227	0.5	1	0	0.815
SVM	ELSY	0.7227	0.5	1	0	0.7922
SVM	GardnerellaVaginallis	0.8434	0.7880	0.9121	0.664	0.779
SVM	GasseriCq	0.7227	0.5	1	0	0.7884
SVM	HPV	0.7227	0.5	1	0	0.7954
SVM	HRHPVCOMPLETE	0.7227	0.5	1	0	0.7884
SVM	HSV12	0.7148	0.4944	0.9888	0	0.7896
SVM	InersCq	0.7227	0.5	1	0	0.7852
SVM	Jensenii	0.7227	0.5	1	0	0.7964
SVM	JenseniiCq	0.7227	0.5	1	0	0.787
SVM	LGasseri	0.7227	0.5	1	0	0.7966
SVM	LIners	0.7227	0.5	1	0	0.79
SVM	LRHPVCOMPLETE	0.7227	0.5	1	0	0.7862
SVM	MegaespheraPhylotipo1	0.8732	0.7918	0.9756	0.608	0.7802
SVM	MG	0.7170	0.4959	0.9919	0	0.7814
SVM	MH	0.8204	0.6894	0.9839	0.395	0.7832
SVM	MultipleHPVCOMPLETE	0.7227	0.5	1	0	0.7804
SVM	MY0911_A	0.7227	0.5	1	0	0.7932
SVM	NG	0.7227	0.5	1	0	0.8014
SVM	pHRHPVCOMPLETE	0.7227	0.5	1	0	0.7748
SVM	RMY0911ELSY	0.7227	0.5	1	0	0.7968
SVM	SingleHPVCOMPLETE	0.7227	0.5	1	0	0.7804
SVM	UP	0.7227	0.5	1	0	0.779
SVM	UU	0.7227	0.5	1	0	0.7976

En los experimentos con SVM utilizando un atributo a la vez se obtuvieron altos niveles de desempeño con atributos tales como "BVCombination", "Atopobium" y "MegaespheraPhylotipo1" con precisiones de 100%, 95% y 87% respectivamente. Además de estos atributos mencionados, se consideran relevantes aquellos que por la precisión promedio

obtenida superan la media aritmética (0.75): “BVNumero”, “GardnerellaVaginallis” y “MH”. Con respecto al tiempo de clasificación, los máximos y mínimos rondan entre los 0.7723 y 0.8682 microsegundos.

Tabla 34. Rendimiento de clasificación promedio de 30 corridas del método *regresión logística* (RL) en el escenario dos al utilizar un atributo del conjunto de datos de la vaginosis bacteriana (VB) a la vez. Los resultados se encuentran ordenados alfabéticamente respecto a la columna “Atributo”. Ms: Microsegundos

Clasificador	Atributos	Precisión	Precisión balanceada	Especificidad	Sensibilidad	Tiempo
RL	Atopobium	0.9534	0.9480	0.9601	0.936	0.7646
RL	BVCombination	**1**	**1**	**1**	**1**	0.8162
RL	BVNumero	0.8243	0.8039	0.8479	0.76	0.7626
RL	Citolog	0.7227	0.5	1	0	0.7772
RL	CitologiaBICAT	0.7227	0.5	1	0	0.7568
RL	CitologiaOrd	0.7227	0.5	1	0	0.7656
RL	Crispatus	0.7227	0.5	1	0	0.7928
RL	CrispatusCq	0.7227	0.5	1	0	0.7598
RL	CT	0.7400	0.53896154	0.9919	0.086	0.7742
RL	EDAD30	0.7227	0.5	1	0	0.8118
RL	EDADENA	0.7227	0.5	1	0	1.111
RL	ELSY	0.7227	0.5	1	0	0.7732
RL	GardnerellaVaginallis	0.8434	0.78808974	0.9121	0.664	0.7628
RL	GasseriCq	0.7227	0.5	1	0	0.7694
RL	HPV	0.7227	0.5	1	0	0.781
RL	HRHPVCOMPLETE	0.7227	0.5	1	0	0.7562
RL	HSV12	0.7148	0.4956	0.9873	0.004	0.7698
RL	InersCq	0.7227	0.5	1	0	0.7724
RL	Jensenii	0.7227	0.5	1	0	0.7708
RL	JenseniiCq	0.7227	0.5	1	0	0.7638
RL	LGasseri	0.7227	0.5	1	0	0.8056
RL	LIners	0.7227	0.5	1	0	0.7654
RL	LRHPVCOMPLETE	0.7227	0.5	1	0	0.7584
RL	MegaespheraPhylotipo1	0.8732	0.7918	0.9756	0.608	0.7704
RL	MG	0.7170	0.4959	0.9919	0	0.7716
RL	MH	0.8204	0.6894	0.9839	0.395	0.7628
RL	MultipleHPVCOMPLETE	0.7227	0.5	1	0	0.7716
RL	MY0911_A	0.7227	0.5	1	0	0.7548
RL	NG	0.7227	0.5	1	0	0.7776
RL	pHRHPVCOMPLETE	0.7227	0.5	1	0	0.7738
RL	RMY0911ELSY	0.7227	0.5	1	0	0.7738
RL	SingleHPVCOMPLETE	0.7227	0.5	1	0	0.7882
RL	UP	0.7227	0.5	1	0	0.7712
RL	UU	0.7227	0.5	1	0	0.7586

De acuerdo a estos resultados, seis atributos de la vaginosis bacteriana ("*BVCombination*", "*Atopobium*", "*MegaespheraPhylotipo1*", "*GardnerellaVaginallis*", "*BVNumero*" y "*MH*") utilizados individualmente como modelos predictivos obtuvieron precisiones promedio por encima de la media (0.75). Se destaca "*BVCombination*" que obtuvo niveles predictivos de 100% en todas las métricas.

Tabla 35. Rendimiento de clasificación promedio de 30 corridas del método *decision tree* (DT) en el escenario dos al utilizar un atributo del conjunto de datos de la vaginosis bacteriana (VB) a la vez. Los resultados se encuentran ordenados alfabéticamente respecto a la columna "Atributo". Ms: Microsegundos.

Clasificador	Atributo	Precisión	Precisión balanceada	Sensibilidad	Especificidad	Tiempo (Ms)
DT	Atopobium	0.9534	0.9480	0.9601	0.936	7.502
DT	BVCombination	**1**	**1**	**1**	**1**	7.483
DT	BVNumero	**1**	**1**	**1**	**1**	7.5252
DT	Citolog	0.7227	0.5	1	0	7.4682
DT	CitologiaBICAT	0.7227	0.5	1	0	7.2974
DT	CitologiaOrd	0.7227	0.5	1	0	7.543
DT	Crispatus	0.7227	0.5	1	0	7.3884
DT	CrispatusCq	0.7227	0.5	1	0	7.6086
DT	CT	0.7227	0.5	1	0	7.5556
DT	EDAD30	0.7227	0.5	1	0	7.4148
DT	EDADENA	0.7227	0.5	1	0	7.2944
DT	ELSY	0.7227	0.5	1	0	7.6586
DT	GardnerellaVaginallis	0.8434	0.7880	0.9121	0.664	7.4506
DT	GasseriCq	0.7134	0.4935	0.9870	0	7.2302
DT	HPV	0.7227	0.5	1	0	7.2522
DT	HRHPVCOMPLETE	0.7227	0.5	1	0	7.4636
DT	HSV12	0.7227	0.5	1	0	7.6942
DT	InersCq	0.7227	0.5	1	0	7.3438
DT	Jensenii	0.7227	0.5	1	0	7.5552
DT	JenseniiCq	0.7227	0.5	1	0	7.6542
DT	LGasseri	0.7227	0.5	1	0	7.5448
DT	LIners	0.7227	0.5	1	0	7.2618
DT	LRHPVCOMPLETE	0.7227	0.5	1	0	7.3812
DT	MegaespheraPhylotipo1	0.873	0.7918	0.9756	0.608	7.563
DT	MG	0.7227	0.5	1	0	8.235
DT	MH	0.8204	0.6894	0.9839	0.395	7.2692
DT	MultipleHPVCOMPLETE	0.7227	0.5	1	0	7.373
DT	MY0911_A	0.7227	0.5	1	0	7.7286
DT	NG	0.7227	0.5	1	0	7.2266
DT	pHRHPVCOMPLETE	0.7227	0.5	1	0	7.2526
DT	RMY0911ELSY	0.7227	0.5	1	0	7.451
DT	SingleHPVCOMPLETE	0.7227	0.5	1	0	7.2422
DT	UP	0.7227	0.5	1	0	9.8296
DT	UU	0.7227	0.5	1	0	8.6704

Los resultados muestran que, mediante la creación de modelos predictivos con árboles de decisión como clasificador utilizando atributos de la vaginosis bacteriana de manera individual, se obtuvieron desempeños por encima del 82% con atributos como "*BVCombination*", "*BVNumero*", "*Atopobium*", "*MegaespheraPhylotipo1*", "*GardnerellaVaginallis*" y "*MH*". De éstos, los dos primeros atributos obtuvieron un desempeño de 100% en todas las métricas calculadas.

Tabla 36. Rendimiento de clasificación promedio de 30 corridas del método *random forest* (RF) en el escenario dos al utilizar un atributo del conjunto de datos de la vaginosis bacteriana (VB) a la vez como modelo predictivo. Los resultados se encuentran ordenados alfabéticamente respecto a la columna "Atributo". Ms: microsegundos.

Clasificador	Atributo	Precisión	Precisión balanceada	Sensibilidad	Especificidad	Tiempo (Ms)
RF	Atopobium	0.9534	0.948	0.960	0.935	1.054
RF	BVCombination	**1**	**1**	**1**	**1**	1.038
RF	BVNumero	**1**	**1**	**1**	**1**	1.091
RF	Citolog	0.7227	0.500	1	0	1.120
RF	CitologiaBICAT	0.7227	0.500	1	0	1.097
RF	CitologiaOrd	0.7227	0.500	1	0	1.166
RF	Crispatus	0.7227	0.500	1	0	1.042
RF	CrispatusCq	0.6481	0.534	0.792	0.275	1.439
RF	CT	0.6481	0.534	0.792	0.275	1.442
RF	EDAD30	0.7227	0.500	1	0	1.047
RF	EDADENA	0.6641	0.488	0.886	0.090	1.388
RF	ELSY	0.6481	0.534	0.792	0.275	1.532
RF	GardnerellaVaginallis	0.8431	0.789	0.913	0.665	1.111
RF	GasseriCq	0.6353	0.464	0.849	0.080	2.562
RF	HPV	0.6481	0.534	0.792	0.275	1.862
RF	HRHPVCOMPLETE	0.6481	0.534	0.792	0.275	1.324
RF	HSV12	0.6481	0.534	0.792	0.275	1.864
RF	InersCq	0.5259	0.409	0.673	0.145	1.582
RF	Jensenii	0.7227	0.500	1	0	1.042
RF	JenseniiCq	0.6298	0.471	0.832	0.110	2.036
RF	LGasseri	0.7227	0.500	1	0	1.013
RF	LIners	0.7227	0.500	1	0	1.015
RF	LRHPVCOMPLETE	0.7227	0.500	1	0	1.413
RF	MegaespheraPhylotipo1	0.8738	0.793	0.976	0.610	1.117
RF	MG	0.6481	0.534	0.792	0.275	1.862
RF	MH	0.8181	0.534	0.792	0.275	1.762
RF	MultipleHPVCOMPLETE	0.6481	0.534	0.792	0.275	1.842
RF	MY0911_A	0.6481	0.534	0.792	0.275	1.762
RF	NG	0.6481	0.534	0.792	0.275	1.662
RF	pHRHPVCOMPLETE	0.6481	0.534	0.792	0.275	1.759
RF	RMY0911ELSY	0.6481	0.534	0.792	0.275	1.486
RF	SingleHPVCOMPLETE	0.6481	0.534	0.792	0.275	1.785
RF	UP	0.6481	0.534	0.792	0.275	1.486
RF	UU	0.6481	0.534	0.792	0.275	1.799

En experimentos con RF, donde se crearon modelos predictivos conformados por los atributos de la vaginosis bacteriana de manera individual, se obtuvieron resultados diversos. 15 atributos obtuvieron índices de precisión por encima de la media (0.71), de los cuales solamente seis obtuvieron precisiones por encima del 80%, y solo dos (“*BVCombination*” y ”*BVNumero*”) obtuvieron niveles predictivos del 100% en todas las métricas calculadas. Respecto al tiempo de clasificación, los modelos experimentados obtuvieron entre 1.013 y 2.562 microsegundos.

Un resumen gráfico de todos los resultados obtenidos de los 10 atributos con mayor precisión obtenido en el escenario dos se proporciona en la Figura 32.

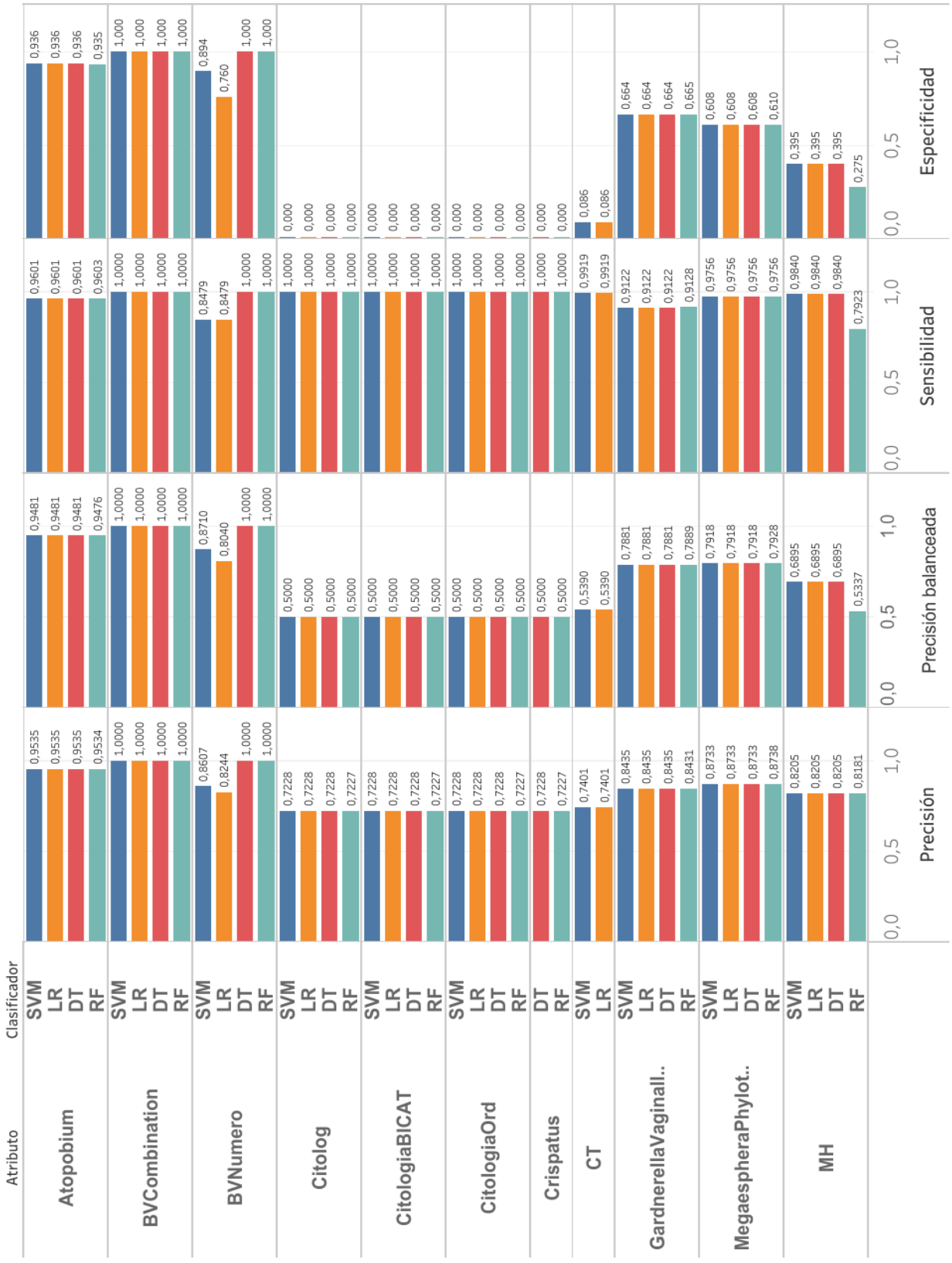

Figura 32. Niveles obtenidos de las métricas experimentadas en el escenario dos de los modelos predictivos de la vaginosis bacteriana. Solamente se muestran los resultados de los primeros 10 atributos con mayor rendimiento obtenido

5.3 Escenario experimental tres

Este escenario se dividió en dos fases experimentales. Ambas fases se basan en la corrida de los clasificadores usando subconjuntos de atributos a partir de la adición incremental de los mismos de igual forma que funcionan los métodos *SFFS* y *SBFS* descritas en la sección 4.2. Para los experimentos se utilizó el orden de importancia de los atributos calculados en la distribución de frecuencias. La creación de los subconjuntos inicia con un solo atributo y finaliza con los 34 que conforman el conjunto de datos de la VB. A continuación, se detallan los experimentos realizados.

1) Modelo predictivo incremental hacia adelante

Cada experimento consiste de 30 corridas de cada método clasificador usando VC-10 al igual que en los escenarios anteriores. A diferencia de estos, en cada corrida se utilizaron subconjuntos de atributos a partir del segundo ranking creado con la distribución de frecuencias. Para la creación de los subconjuntos se implementó la idea general del método *SFFS* descrito en la sección 4.2. Es decir, el espacio de atributos se incrementó de manera gradual agregando un atributo a la vez: se inició con el atributo más relevante y se incrementó hasta finalizar con el menos relevante de acuerdo al ranking calculado con la distribución de frecuencias (Sección 4.4). Tal cual como en el escenario uno y dos, en este escenario se calcularon y promediaron las métricas definidas en la sección 2.4 con base en las 30 corridas. Las Tablas 37, 38, 39 y 40 muestran los resultados obtenidos por SVM, RL, DT y RF, respectivamente.

Tabla 37. Rendimiento promedio de las 30 corridas de los modelos predictivos de la vaginosis bacteriana (VB) con *máquina de vector soporte* (*SVM*) como método clasificador al utilizar los atributos de manera incremental hacia adelante.

Clasificador	No. Atributos	Precisión	Precisión balanceada	Sensibilidad	Especificidad	Tiempo (Ms)
SVM	1	1	1	1	1	0.7411
SVM	2	1	1	1	1	0.7716
SVM	3	1	1	1	1	0.8285
SVM	4	1	1	1	1	0.8663
SVM	5	1	1	1	1	0.8533
SVM	6	1	1	1	1	0.8236
SVM	7	1	1	1	1	0.9147
SVM	8	1	1	1	1	1.0512
SVM	9	1	1	1	1	0.9683
SVM	10	1	1	1	1	0.9328
SVM	11	1	1	1	1	0.8404
SVM	12	1	1	1	1	1.0022
SVM	13	1	1	1	1	0.8703
SVM	14	1	1	1	1	0.8992
SVM	15	1	1	1	1	0.8331
SVM	16	1	1	1	1	0.8016
SVM	17	1	1	1	1	0.8482
SVM	18	1	1	1	1	1.0459
SVM	19	1	1	1	1	0.832
SVM	20	1	1	1	1	0.9563
SVM	21	1	1	1	1	0.9793
SVM	22	1	1	1	1	0.9179
SVM	23	1	1	1	1	0.876
SVM	24	1	1	1	1	0.8833
SVM	25	1	1	1	1	0.8874
SVM	26	1	1	1	1	0.8916
SVM	27	1	1	1	1	0.8862
SVM	28	1	1	1	1	0.8871
SVM	29	1	1	1	1	0.9129
SVM	30	1	1	1	1	0.8998
SVM	31	1	1	1	1	0.9461
SVM	32	1	1	1	1	0.9375
SVM	33	1	1	1	1	0.971
SVM	34	1	1	1	1	1.1472

En estos experimentos, SVM obtuvo el 100% de predicciones correctamente clasificadas en todos los casos, aun con la utilización de solamente un atributo de la vaginosis bacteriana. Aún con el incremento del número de atributos que conformaban los modelos experimentados, el rendimiento del SVM se mantuvo al margen del 100% en todas las métricas calculadas. El

tiempo de clasificación se incrementó conforme se agregaban más atributos al subconjunto de datos.

Tabla 38. Rendimiento promedio de las 30 corridas de los modelos predictivos de la vaginosis bacteriana (VB) con *regresión logística* (RL) como método clasificador al utilizar los subconjuntos de atributos de manera incremental hacia adelante.

Clasificador	No. Atributos	Precisión	Precisión balanceada	Sensibilidad	Especificidad	Tiempo (Ms)
RL	1	1	1	1	1	0.8108
RL	2	1	1	1	1	0.8061
RL	3	1	1	1	1	0.8473
RL	4	1	1	1	1	0.802
RL	5	1	1	1	1	0.9608
RL	6	1	1	1	1	0.878
RL	7	1	1	1	1	1.049
RL	8	1	1	1	1	1.0077
RL	9	1	1	1	1	0.8728
RL	10	1	1	1	1	0.9351
RL	11	1	1	1	1	1.0577
RL	12	1	1	1	1	0.9656
RL	13	1	1	1	1	0.9513
RL	14	1	1	1	1	0.974
RL	15	1	1	1	1	0.9869
RL	16	1	1	1	1	0.8982
RL	17	1	1	1	1	0.9277
RL	18	1	1	1	1	0.9149
RL	19	1	1	1	1	0.9153
RL	20	1	1	1	1	0.9604
RL	21	1	1	1	1	1.0255
RL	22	1	1	1	1	0.9509
RL	23	1	1	1	1	0.9511
RL	24	1	1	1	1	0.9864
RL	25	1	1	1	1	1.187
RL	26	1	1	1	1	1.135
RL	27	1	1	1	1	1.0315
RL	28	1	1	1	1	1.1528
RL	29	1	1	1	1	1.0726
RL	30	1	1	1	1	1.1228
RL	31	1	1	1	1	1.0528
RL	32	1	1	1	1	1.1009
RL	33	1	1	1	1	1.1117
RL	34	1	1	1	1	1.0842

La creación de modelos predictivos con RL utilizando los atributos de manera incremental obtuvo índices del 100% en las métricas de desempeño como precisión, precisión balanceada,

sensibilidad y especificidad. El tiempo de clasificación aumentó de igual manera, incrementándose conforme se agregaban más atributos al modelo.

Tabla 39. Rendimiento promedio de las 30 corridas de los modelos predictivos de la vaginosis bacteriana (VB) con *árboles de decisión* (DT) al utilizar los subconjuntos de manera incremental hacia adelante.

Clasificador	No. atributos	Precisión	Precisión balanceada	Sensibilidad	Especificidad	Tiempo (Ms)
DT	1	1	1	1	1	8.1868
DT	2	1	1	1	1	12.1382
DT	3	1	1	1	1	9.3722
DT	4	1	1	1	1	9.5546
DT	5	1	1	1	1	8.4558
DT	6	1	1	1	1	13.9602
DT	7	1	1	1	1	14.8442
DT	8	1	1	1	1	19.5524
DT	9	1	1	1	1	13.4584
DT	10	1	1	1	1	8.9272
DT	11	1	1	1	1	9.0088
DT	12	1	1	1	1	9.2408
DT	13	1	1	1	1	8.9596
DT	14	1	1	1	1	9.7874
DT	15	1	1	1	1	10.3374
DT	16	1	1	1	1	9.5108
DT	17	1	1	1	1	9.99
DT	18	1	1	1	1	10.2968
DT	19	1	1	1	1	10.4586
DT	20	1	1	1	1	10.0156
DT	21	1	1	1	1	10.2126
DT	22	1	1	1	1	11.158
DT	23	1	1	1	1	10.5928
DT	24	1	1	1	1	10.7642
DT	25	1	1	1	1	11.2454
DT	26	1	1	1	1	10.9716
DT	27	1	1	1	1	11.9408
DT	28	1	1	1	1	12.5888
DT	29	1	1	1	1	11.603
DT	30	1	1	1	1	12.8586
DT	31	1	1	1	1	12.8532
DT	32	1	1	1	1	12.2708
DT	33	1	1	1	1	11.9022
DT	34	1	1	1	1	14.3762

El nivel de desempeño de DT fue exactamente igual en todos los experimentos, tanto al iniciar con solamente el atributo más relevante de la VB como al utilizar los 34 atributos. Los mejores

tiempos de clasificación fueron obtenidos al utilizar dos, uno y nueve de los atributos más relevantes de la VB, respectivamente.

Tabla 40. Rendimiento promedio de las 30 corridas de los modelos predictivos de la vaginosis bacteriana (VB) con *bosques aleatorios* (RF) al utilizar los subconjuntos de atributos agregados de manera incremental hacia adelante

Clasificador	No. Atributos	Precisión	Precisión balanceada	Sensibilidad	Especificidad	Tiempo (Ms)
RF	1	1	1	1	1	2.4294
RF	2	1	1	1	1	1.5914
RF	3	1	1	1	1	2.6378
RF	4	1	1	1	1	3.3232
RF	5	1	1	1	1	3.1088
RF	6	1	1	1	1	3.3658
RF	7	1	1	1	1	2.6382
RF	8	1	1	1	1	2.7602
RF	9	1	1	1	1	2.6174
RF	10	1	1	1	1	2.7386
RF	11	1	1	1	1	2.836
RF	12	1	1	1	1	3.2146
RF	13	1	1	1	1	3.031
RF	14	1	1	1	1	3.1286
RF	15	1	1	1	1	3.2582
RF	16	1	1	1	1	3.3932
RF	17	1	1	1	1	3.473
RF	18	1	1	1	1	3.5794
RF	19	1	1	1	1	3.681
RF	20	1	1	1	1	3.9544
RF	21	1	1	1	1	3.9078
RF	22	1	1	1	1	4.3534
RF	23	1	1	1	1	4.128
RF	24	1	1	1	1	4.275
RF	25	1	1	1	1	4.3784
RF	26	1	1	1	1	4.4948
RF	27	1	1	1	1	4.5734
RF	28	1	1	1	1	4.6736
RF	29	1	1	1	1	4.8012
RF	30	1	1	1	1	4.9366
RF	31	1	1	1	1	4.9894
RF	32	1	1	1	1	5.1072
RF	33	1	1	1	1	5.262
RF	34	1	1	1	1	5.2576

El desempeño de los modelos predictivos con RF fue óptimo en todos los experimentos al obtener el 100% de precisión como promedio de las corridas realizadas. Este nivel se obtuvo a partir de los modelos creados con únicamente el atributo más relevante de la VB y se mantuvo

hasta aumentar a los 34. Los resultados en cuanto a las métricas de precisión balanceada, sensibilidad y especificidad fueron exactamente iguales. El tiempo de clasificación más bajo fue de 1.59 microsegundos y el más alto de 5.26 microsegundos, habiendo aumentado de manera incrementa conforme aumentaba el número de atributos del modelo experimentado.

Un resumen gráfico de todos los resultados obtenidos en los experimentos de los modelos predictivos incrementales hacia adelante se presenta en la Figura 33.

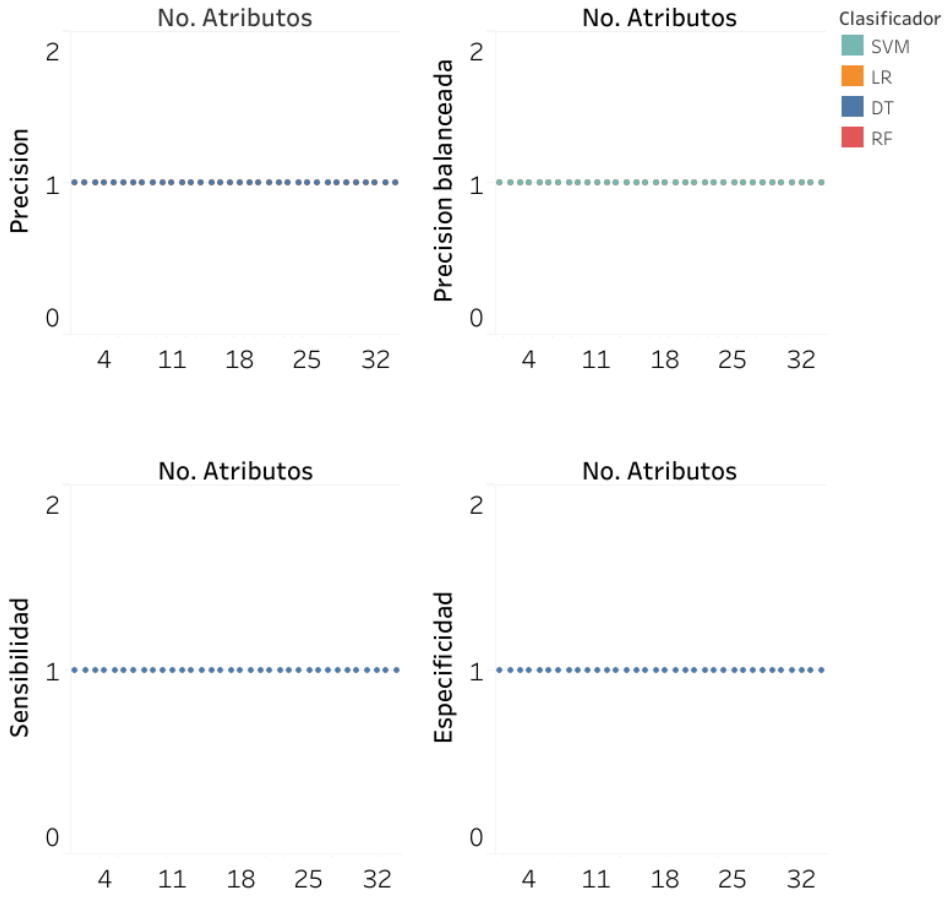

Figura 33. Gráfico comparativo del rendimiento de los clasificadores al utilizar los subconjuntos de atributos de manera incremental hacia adelante. DT: decision trees, RF: random forests, RL: regresión logística, SVM: support vector machine.

Como se muestra en los resultados, todos los clasificadores utilizando un subconjunto creado con el atributo más relevante de la vaginosis bacteriana obtuvieron un rendimiento de 100% en las métricas calculadas (precisión, precisión balanceada, especificidad y sensibilidad). Es decir, a partir de la utilización del atributo "*Atopobium*" en simultaneidad con la clase principal como modelo predictivo de la vaginosis bacteriana, se obtuvo un 100% en la precisión, precisión balanceada, especificidad y sensibilidad en los experimentos de todos los clasificadores. Conforme el número de atributos aumentaba, las métricas se mantuvieron al 100%.

Respecto a los tiempos de clasificación obtenidos por los experimentados en esta fase, la Figura 34 muestra los resultados calculados en microsegundos.

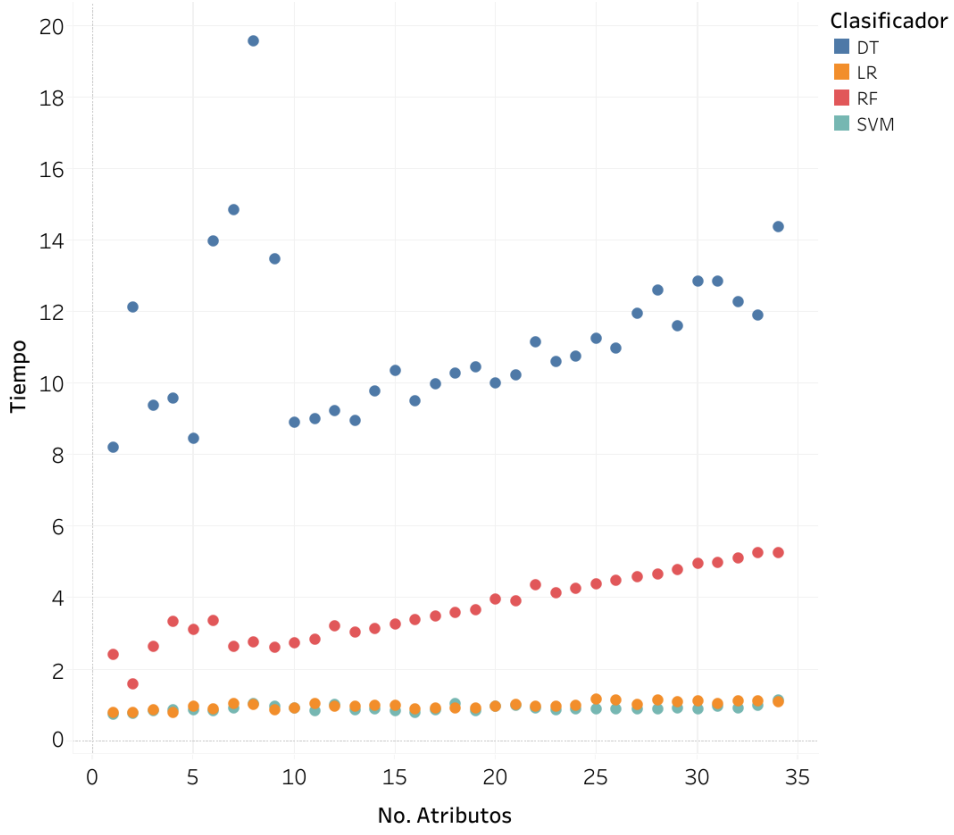

Figura 34.Tiempo de clasificación obtenido por los métodos clasificadores en los experimentos con los modelos predictivos incrementales hacia adelante

2) Modelo predictivo incremental hacia atrás

Los experimentos consisten en la corrida de los clasificadores definidos en el capítulo 2.3 utilizando la idea general del método *SBFS*, al utilizar subconjuntos de atributos agregados de manera incremental hacia atrás. Es decir, el espacio de atributos se incrementó de manera gradual agregando un atributo a la vez iniciando con el atributo menos relevante hasta finalizar con el más relevante de acuerdo al ranking calculado con la distribución de frecuencias (Sección 4.4). De la misma forma que los experimentos anteriores, se obtuvieron los promedios de las 30 corridas bajo el esquema de validación cruzada de las métricas definidas en la Sección 2.5. Los resultados se muestran en la Tabla 41.

Tabla 41. Rendimiento promedio de las 30 corridas de *máquina de vector soporte* (SVM) al utilizar los atributos de la vaginosis bacteriana de manera incremental hacia atrás a partir del segundo ranking general de atributos.

Clasificador	No. atributos	Precisión	Precisión balanceada	Sensibilidad	Especificidad	Tiempo (Ms)
SVM	1	0.7228	0.5000	1	0	0.7630
SVM	2	0.7228	0.5000	1	0	0.7570
SVM	3	0.7228	0.5000	1	0	0.7527
SVM	4	0.7209	0.4987	0.9974	0	0.7490
SVM	5	0.7209	0.4987	0.9974	0	0.7613
SVM	6	0.7228	0.5000	1	0	0.7550
SVM	7	0.7209	0.4987	0.9974	0	0.7597
SVM	8	0.7190	0.4973	0.9947	0	0.7687
SVM	9	0.7114	0.4921	0.9842	0	0.7750
SVM	10	0.7055	0.4880	0.9761	0	0.8803
SVM	11	0.7037	0.4868	0.9735	0	1.1117
SVM	12	0.7057	0.4880	0.9761	0	1.0233
SVM	13	0.7057	0.4881	0.9763	0	0.9150
SVM	14	0.7036	0.4866	0.9733	0	0.9727
SVM	15	0.7092	0.4906	0.9812	0	0.8380
SVM	16	0.7037	0.5026	0.9519	0.0533	0.8297
SVM	17	0.7055	0.5039	0.9545	0.0533	0.8320
SVM	18	0.7055	0.5039	0.9545	0.0533	1.0570
SVM	19	0.7054	0.5038	0.9543	0.0533	1.0237
SVM	20	0.7054	0.5038	0.9543	0.0533	1.9053
SVM	21	0.7037	0.5026	0.9519	0.0533	1.8657
SVM	22	0.7074	0.5052	0.9571	0.0533	2.9650
SVM	23	0.7247	0.5329	0.9624	0.1033	2.9820
SVM	24	0.7228	0.5335	0.9571	0.11	6.4613
SVM	25	0.7209	0.5322	0.9545	0.11	4.9690
SVM	26	0.7151	0.5242	0.9517	0.0967	4.6883
SVM	27	0.7060	0.5166	0.9415	0.0917	3.7137

Tabla 41. (Continuación)

Clasificador	No. atributos	Precisión	Precisión balanceada	Sensibilidad	Especificidad	Tiempo (Ms)
SVM	28	0.7021	0.5111	0.9389	0.0833	2.4460
SVM	29	0.7992	0.6761	0.9521	0.4	2.5593
SVM	30	0.8708	0.8329	0.9175	0.7483	2.0160
SVM	31	0.9206	0.8999	0.9464	0.8533	1.6313
SVM	32	0.9247	0.9007	0.9547	0.8467	1.3830
SVM	33	1	1	1	1	1.7990
SVM	34	1	1	1	1	1.1857

Los resultados muestran que SVM obtuvo niveles de rendimiento por debajo del 79% de capacidad predictiva usando los primeros 29 atributos menos relevantes. No fue, sino a partir de los experimentos realizados con modelos conformados por los 30 atributos menos relevantes que se obtuvieron métricas de desempeño por encima del 87% de precisión. Los niveles de precisión balanceada se incrementaron de manera similar a la precisión. Los índices de sensibilidad fueron cercanos al 100% a partir de los primeros experimentos, caso contrario a los índices de especificidad, que fueron incrementados gradualmente conforme aumentaba el espacio de atributos. Se obtuvieron tiempos de clasificación desde 0.74 hasta 6.46 microsegundos.

Los experimentos con el método de *regresión logística* se completaron, y los resultados se muestran en la Tabla 42.

Tabla 42. Rendimiento promedio de las 30 corridas de *regresión logística* (RL) al agregar al modelo los atributos de la vaginosis bacteriana de manera incremental hacia atrás.

Clasificador	No. atributos	Precisión	Precisión balanceada	Sensibilidad	Especificidad	Tiempo (Ms)
RL	1	0.7228	0.5	1	0	0.777
RL	2	0.7172	0.4962	0.9923	0	0.825
RL	3	0.7172	0.4962	0.9923	0	0.919
RL	4	0.7172	0.4962	0.9923	0	1.056
RL	5	0.7116	0.4923	0.9846	0	2.426
RL	6	0.7060	0.4885	0.9769	0	0.957
RL	7	0.7116	0.4985	0.9769	0.02	0.976
RL	8	0.7057	0.4943	0.9686	0.02	0.945
RL	9	0.7112	0.5341	0.9282	0.14	0.922
RL	10	0.7054	0.5241	0.9282	0.12	0.938
RL	11	0.7054	0.5241	0.9282	0.12	0.937
RL	12	0.7057	0.5241	0.9282	0.12	1.339
RL	13	0.7116	0.5283	0.9365	0.12	1.885
RL	14	0.7112	0.5279	0.9359	0.12	2.101
RL	15	0.6994	0.5199	0.9199	0.12	1.517
RL	16	0.6994	0.5199	0.9199	0.12	3.154
RL	17	0.6994	0.5199	0.9199	0.12	1.158
RL	18	0.6883	0.5061	0.9122	0.1	1.782
RL	19	0.6824	0.5019	0.9038	0.1	1.882
RL	20	0.6824	0.5019	0.9038	0.1	1.723
RL	21	0.6883	0.5061	0.9122	0.1	1.584
RL	22	0.6707	0.4878	0.8955	0.08	1.143
RL	23	0.6991	0.5378	0.8955	0.18	1.577
RL	24	0.7050	0.5419	0.9038	0.18	1.09
RL	25	0.6994	0.5319	0.9038	0.16	1.105
RL	26	0.6825	0.5404	0.8558	0.225	0.988
RL	27	0.6763	0.5359	0.8468	0.225	1.029
RL	28	0.6638	0.5377	0.8154	0.26	1.147
RL	29	0.7567	0.6671	0.8641	0.47	1.259
RL	30	0.8433	0.7981	0.8962	0.7	1.033
RL	31	0.9186	0.8971	0.9442	0.85	1.046
RL	32	1	1	1	1	1.18
RL	33	1	1	1	1	2.285
RL	34	1	1	1	1	1.139

Los niveles de rendimiento obtenidos por RL en estos experimentos se incrementaron considerablemente a partir de la implementación de modelos predictivos creados con los primeros 29 atributos menos relevantes. No fue, sino hasta la utilización de los primeros 30 atributos menos relevantes que se obtuvieron rendimientos por encima del 80% de precisión.

La especificidad y sensibilidad obtenida por RL de igual manera incrementaron conforme aumentó el número de atributos agregados al modelo. El tiempo de clasificación fue variado, obteniendo resultados desde 0.777 hasta 3.154 microsegundos.

A su vez, los resultados obtenidos por *árboles de decisión* se muestran en la Tabla 43.

Tabla 43. Rendimiento promedio de las 30 corridas de *decision tree* (DT) al agregar al modelo los atributos de la vaginosis bacteriana de manera incremental hacia atrás

Clasificador	No. Atributos	Precisión	Precisión balanceada	Sensibilidad	Especificidad	Tiempo
DT	1	0.7227	0.5	1	0	8.33
DT	2	0.7227	0.5	1	0	7.504
DT	3	0.7227	0.5	1	0	7.546
DT	4	0.7227	0.5	1	0	8.227
DT	5	0.7227	0.5	1	0	8.594
DT	6	0.7227	0.5	1	0	10.676
DT	7	0.7227	0.5	1	0	10.089
DT	8	0.7109	0.4916	0.9833	0	8.482
DT	9	0.7227	0.5	1	0	9.952
DT	10	0.7227	0.5	1	0	10.68
DT	11	0.7227	0.5	1	0	9.715
DT	12	0.7227	0.5	1	0	10.562
DT	13	0.7227	0.5	1	0	11.833
DT	14	0.7227	0.5	1	0	9.477
DT	15	0.7227	0.5	1	0	9.454
DT	16	0.7227	0.5	1	0	9.63
DT	17	0.7227	0.5	1	0	10.686
DT	18	0.7227	0.5	1	0	9.815
DT	19	0.7227	0.5	1	0	10.114
DT	20	0.7109	0.4919	0.9839	0	10.406
DT	21	0.7171	0.4961	0.9923	0	17.395
DT	22	0.7171	0.4961	0.9923	0	26.642
DT	23	0.7171	0.4961	0.9923	0	21.243
DT	24	0.7171	0.4961	0.9923	0	23.377
DT	25	0.7227	0.5061	0.9923	0.02	19.518
DT	26	0.7227	0.5	1	0	20.331
DT	27	0.7109	0.4981	0.9762	0.02	26.298
DT	28	0.7171	0.4961	0.9923	0	30.446
DT	29	0.8145	0.6859	0.9769	0.395	51.32
DT	30	0.8607	0.8192	0.9134	0.725	38.249
DT	31	0.8735	0.8522	0.9044	0.8	131.56
DT	32	0.9882	0.98	1	0.96	77.34
DT	33	0.9882	0.98	1	0.96	77.34
DT	34	1	1	1	1	117.731

En los experimentos con DT se obtuvieron niveles de precisión por encima del 80% a partir del uso de los primeros 29 atributos menos relevantes de la VB. La precisión balanceada obtuvo un incremento similar, pero no fue sino hasta la utilización de los primeros 30 atributos menos relevantes que superó el 80%. La sensibilidad y especificidad tuvieron comportamientos a la inversa. Es decir, la sensibilidad inició con 100% y se mantuvo a lo largo de los experimentos. La especificidad inicio con 0% en los primeros experimentos e incrementó a partir de la utilización de los primeros 29 atributos. Respecto a los tiempos de clasificación, se obtuvieron tiempos entre 7.504 y 131.56 microsegundos correspondientes a el mínimo y máximo respectivamente.

Finalmente, los resultados de los experimentos con *bosques aleatorios* se muestran en la Tabla 44.

Tabla 44. Rendimiento promedio de las 30 corridas de *decision tree* (DT) al agregar al modelo los atributos de la vaginosis bacteriana de manera incremental hacia atrás

Clasificador	No. Atributos	Precisión	Precisión balanceada	Sensibilidad	Especificidad	Tiempo
RF	1	0.72271242	0.5	1	0	8.33
RF	2	0.73447712	0.52	1	0.04	2.677
RF	3	0.72826797	0.51	1	0.02	2.728
RF	4	0.72271242	0.5	1	0	2.896
RF	5	0.73447712	0.52	1	0.04	2.642
RF	6	0.73447712	0.52	1	0.04	2.677
RF	7	0.71715686	0.502307692	0.98461538	0.02	2.512
RF	8	0.72826797	0.51	1	0.02	2.728
RF	9	0.71058007	0.503333333	0.96666667	0.04	3.103
RF	10	0.71683007	0.501987179	0.98397436	0.02	3.276
RF	11	0.71683007	0.510320513	0.97564103	0.045	3.542
RF	12	0.71160131	0.492307692	0.98461538	0	4.422
RF	13	0.72271242	0.5	1	0	4.938
RF	14	0.72271242	0.5	1	0	4.851
RF	15	0.72271242	0.5	1	0	7.046
RF	16	0.72271242	0.5	1	0	7.04
RF	17	0.72271242	0.5	1	0	9.539
RF	18	0.72271242	0.5	1	0	15.14
RF	19	0.72271242	0.5	1	0	47.046
RF	20	0.72271242	0.5	1	0	38.648
RF	21	0.71683007	0.495833333	0.99166667	0	31.391
RF	22	0.71683007	0.495833333	0.99166667	0	10.966
RF	23	0.71058007	0.491666667	0.98333333	0	20.759

Tabla 44. (Continuación)

Clasificador	No. atributos	Precisión	Precisión balanceada	Sensibilidad	Especificidad	Tiempo (Ms)
RF	24	0.70502451	0.487820513	0.97564103	0	14.081
RF	25	0.71683007	0.495833333	0.99166667	0	75.199
RF	26	0.71683007	0.495833333	0.99166667	0	28.109
RF	27	0.70539216	0.487820513	0.97564103	0	32.653
RF	28	0.70539216	0.487820513	0.97564103	0	56.383
RF	29	0.71683007	0.495833333	0.99166667	0	56.264
RF	30	0.75539216	0.639423077	0.90384615	0.375	216.711
RF	31	0.89464869	0.863910256	0.93782051	0.79	56.651
RF	32	0.91895425	0.914423077	0.92884615	0.9	81.838
RF	33	1	1	1	1	119.616
RF	34	1	1	1	1	5.7052

En resumen, el comportamiento de RF al utilizar subconjuntos de atributos de manera secuencial hacia atrás fue de la siguiente manera. Los modelos predictivos creados con los primeros 30 atributos menos relevantes de la vaginosis bacteriana obtuvieron niveles de precisión por debajo del 80%. A su vez, la precisión balanceada y especificidad alcanzaron niveles aceptados hasta la utilización de los primeros 31 atributos menos relevantes. NO así con la sensibilidad, que obtuvo niveles aceptables desde los primeros experimentos. Respecto a los tiempos, hubieron tiempos de clasificación entre 2.512 y 216.711 microsegundos.

Un resumen gráfico de todos los resultados obtenidos en los experimentos de los modelos predictivos incrementales hacia adelante se presenta en la Figura 35.

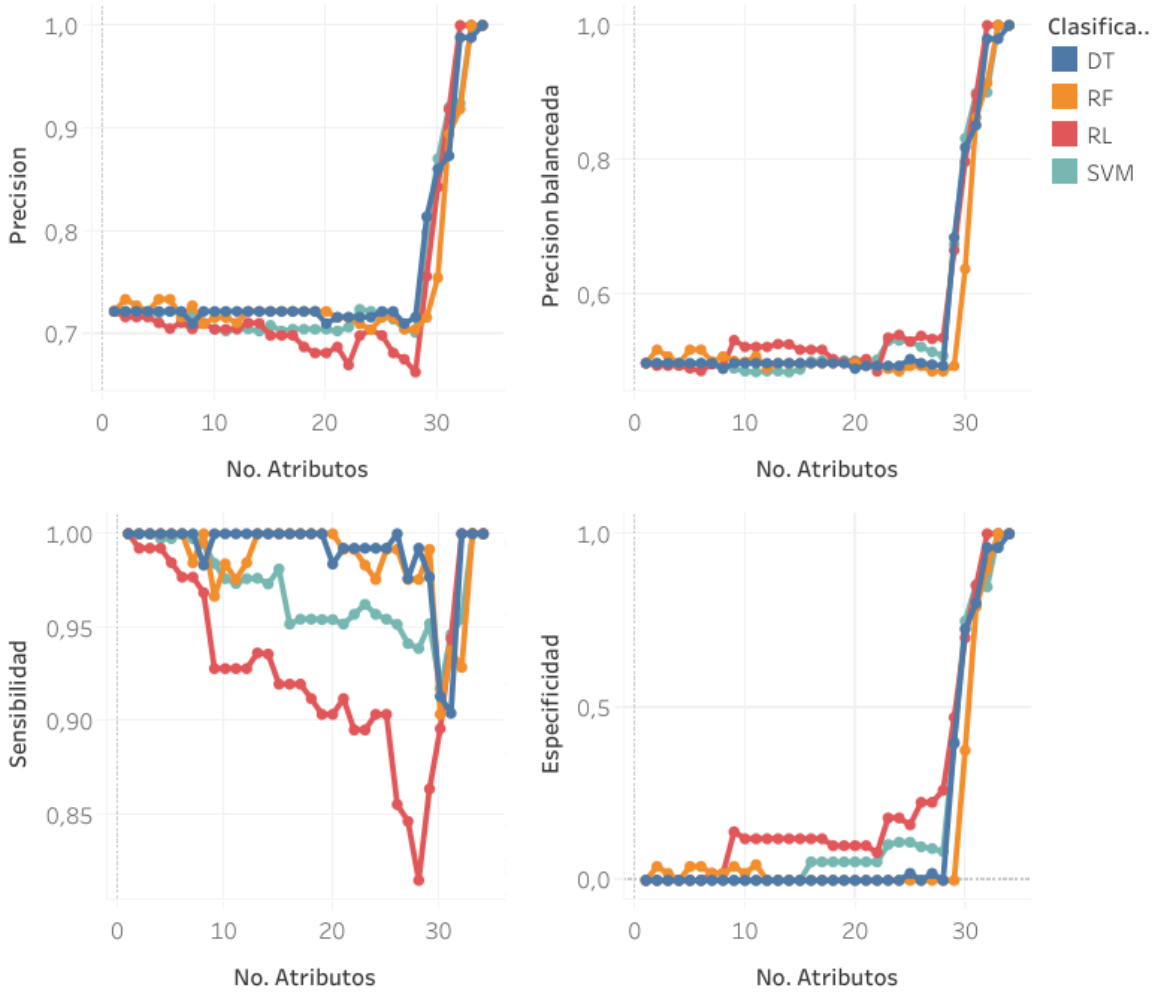

Figura 35. Gráfico comparativo del rendimiento de los clasificadores al utilizar los subconjuntos de atributos de manera incremental hacia atrás. DT: decision trees, RF: random forests, RL: regresión logística, SVM: support vector machine.

Respecto a los tiempos de clasificación obtenidos por los experimentados en esta fase, la Figura 36 muestra los resultados calculados en microsegundos.

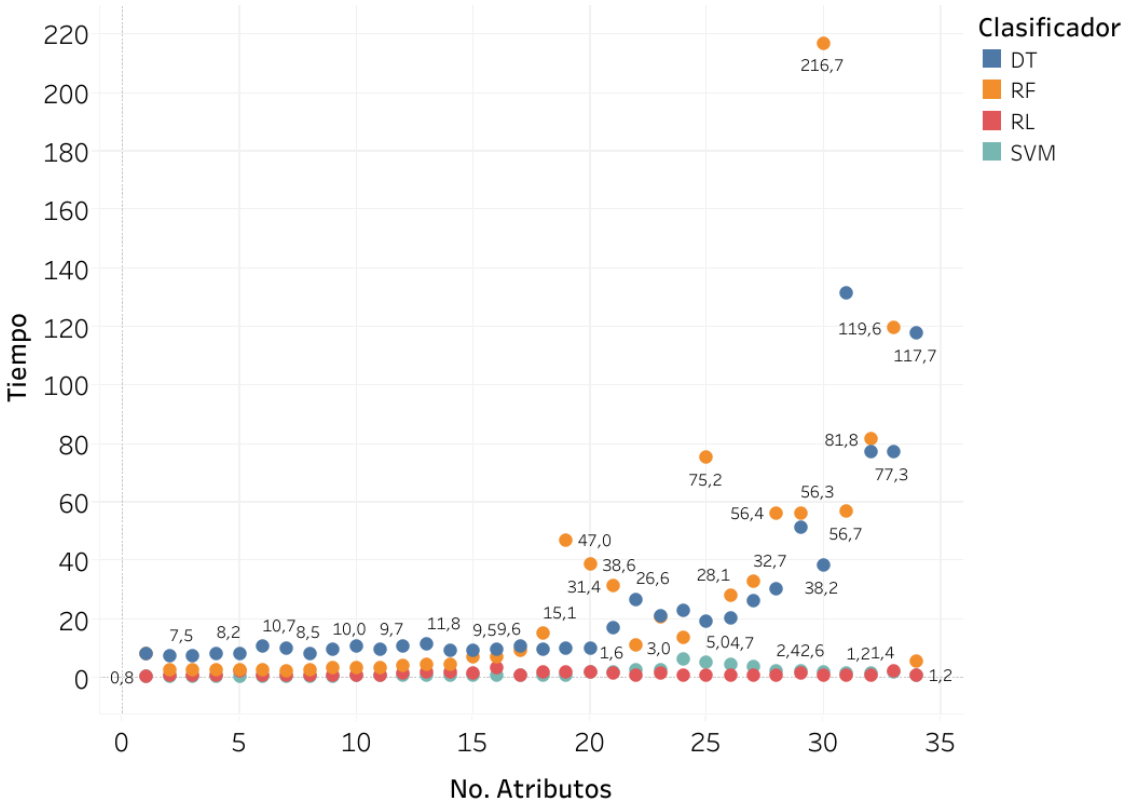

Figura 36. Tiempo -en microsegundos- de clasificación promedio obtenido por los métodos clasificadores en los experimentos con los modelos predictivos incrementales hacia adelante. DT: decision trees, RF: random forests, RL: regresión logística, SVM: support vector machine.

Los resultados muestran que a mayor número de atributos que conforman el subconjunto de datos, más alto es el costo computacional. En este caso, el tiempo de clasificación se ve afectado por el número de atributos que conforman el espacio de atributos. En los experimentos realizados con los clasificadores, es RF el algoritmo que requiere más tiempo para sus procedimientos de clasificación. De los clasificadores “más rápidos” podemos encontrar a SVM y RL.

Capítulo 6. Conclusiones y trabajos futuros.

6.1 Conclusiones

En este trabajo se determinaron los atributos más relevantes para la predicción de la VB a partir de un conjunto de datos microbiológico mediante la implementación de diferentes métodos del área de aprendizaje automático para la clasificación, selección de atributos y reducción de dimensiones. Después de realizar diversos experimentos se obtuvieron dos rankings de atributos que representan los atributos más relevantes para el diagnóstico de la VB. El primero de ellos se realizó con base en el índice de relevancia de los atributos. El segundo se calculó con base en un análisis de frecuencias. Finalmente, se corrieron experimentos con modelos predictivos de la vaginosis bacteriana utilizando como base cuatro distintos métodos clasificadores implementados en tres escenarios distintos. Esto a su vez, permitió identificar el numero óptimo de predictores de la vaginosis bacteriana indispensables. Por efecto de las acciones realizadas y los resultados obtenidos, se consideran cumplidos los objetivos específicos planteados en este proyecto.

Mediante los métodos de selección de atributos se determinó que “BVCombination” es el atributo más relevante entre las características de la vaginosis bacteriana. La importancia de este atributo es que denota la combinación de microorganismos previamente identificados relacionados con la vaginosis bacteriana (bacterias anaerobias tales como *Gardnerella vaginallis, Prevotella spp., Mycoplasma hominis*, entre otras), y representa la presencia o ausencia de dichos microorganismos en la vía vaginal. Por otro lado, al destacar la importancia de las características que representan el recuento de microorganismos en el conjunto de datos, “Atopobium” se considera como la característica más relacionada con el diagnóstico de la VB. Es importante mencionar, como parte de la validación clínica de los hallazgos, que este enfoque resulta esencial para el diagnóstico de la VB porque, en la mayoría de los casos, la transición entre el estado normal y el de enfermedad se basa en la densidad microbiana de los microorganismos que suelen estar presentes en el microambiente cervicovaginal. Desde este

punto de vista biológico, aunque la presencia de *Gardnerella vaginalis* se relaciona comúnmente con el desarrollo de la vaginosis bacteriana, la mayoría de los estudios sugieren que la principal característica que distingue el rol de *Gardnerella* es la alta densidad observada en muestras vaginales, que frecuentemente se asocia con su comportamiento patológico. Por otro lado, aunque la presencia de *Atopobium vaginae* y/o el de *Megasphaera phylotype 1* no se asocian frecuentemente a signos clínicos de VB, los estudios moleculares han demostrado su alta tasa de prevalencia en mujeres con diagnóstico confirmado de VB.

Los diferentes modelos predictivos experimentados con solamente “BVCombination” en combinación con la clase principal confirman que se pueden obtener niveles de precisión del 100%, o cercano a ello, utilizando cualquiera de los cuatro métodos clasificadores implementados en este trabajo. Sin embargo, al haber una mínima diferencia en cuanto a los tiempos de clasificación obtenidos, se determinó que el mejor clasificador para la VB es la máquina de vector soporte, tratada durante todo el proyecto como SVM. Los resultados confirman que SVM es un clasificador altamente preciso tanto en el uso del conjunto de atributos completo como en el uso de solamente un subconjunto de atributos reducido de la vaginosis bacteriana. Desde el punto de vista microbiológico se destaca la utilidad potencial de SVM para identificar aquellos microorganismos relacionados con la etiología de la VB, lo que permitiría reducir el número de ensayos de laboratorios necesarios para determinar la presencia de vaginosis bacteriana con alta precisión diagnóstica.

6.2 Trabajos futuros

Durante el tiempo en que este proyecto fue implementado, surgió nueva información respecto a la composición microbiana de la vaginosis bacteriana, cuyos datos están enfocados en la cuantificación de microorganismos presentes en las muestras cervicales. Resultaría interesante planificar un estudio similar al realizado en este proyecto con la nueva información puramente microbiana.

A lo largo de los diversos experimentos realizados en este proyecto, se obtuvo información suficiente que demuestra que la utilización de técnicas para el balanceo de clases aplicadas en

la fase de preprocesamiento puede mejorar el nivel de precisión de los modelos predictivos. Dichas técnicas no fueron implementadas en este proyecto puesto que su utilidad se alejaba de los objetivos planteados, pero como trabajo futuro se sugieren experimentos con técnicas de balanceo de datos tales como el sobremuestreo o submuestreo para la creación de instancias sintéticas, por mencionar algunos.

Referencias bibliográficas

[1] C. Aggarwal, Data Classification: Algorithms and Applications, Chapman & Hall/CRC Press, Boca Raton, Florida, USA, 2014.

[2] J. Ali, R. Khan, N. Ahmad, I. Maqsood, Random Forests and Decision Trees, Int. J. Comput. Sci. Issues. 9 (2012) 272–278.

[3] J. Atashili, C. Poole, P.M. Ndumbe, A.A. Adimora, J.S. Smith, Bacterial vaginosis and HIV acquisition: A meta-analysis of published studies, Aids. 22 (2008) 1493–1501.

[4] Y.S. Baker, D. Beck, R. Agrawal, G. Dozier, J.A. Foster, Detecting Bacterial Vaginosis using machine learning, in: Proc. 2014 ACM Southeast Reg. Conf. ACM SE 2014, 2014: pp. 1–4.

[5] D. Beck, J.A. Foster, Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics, PLoS One. 9 (2014).

[6] D. Beck, J.A. Foster, Machine learning classifiers provide insight into the relationship between microbial communities and bacterial vaginosis, BioData Min. 8 (2015) 1–9.

[7] P. Belda-Ferre, L.D. Alcaraz, R. Cabrera-Rubio, H. Romero, A. Simón-Soro, M. Pignatelli, A. Mira, The oral metagenome in health and disease, ISME J. 6 (2012) 46–56.

[8] J. Benesty, J. Chen, Y. Huang, I. Cohen, Optimal filters in the time domain, Springer Top. Signal Process. 2 (2009) 1–18.

[9] B. Bischl, M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, Z.M. Jones, Mlr: Machine learning in R, J. Mach. Learn. Res. 17 (2016) 1–5.

[10] M. Bramer, Introduction to Data Mining, in: 2013: pp. 1–8.

[11] M. Bramer, Principles of Data Mining, Third Edit, Springer London, London, 2016.

[12] L. Breiman, Random forests, in: Random For., CRC Press, First. | Boca Raton : CRC Press, 2019., 2001: pp. 1–122.

[13] P. Bühlmann, S. van de Geer, Statistics for High-Dimensional Data: Methods, Theory and Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[14] P.L. Buttigieg, A. Ramette, A guide to statistical analysis in microbial ecology: A community-focused, living review of multivariate data analyses, FEMS Microbiol. Ecol. 90

(2014) 543–550.

[15] J. Canul-Reich, An iterative feature perturbation method for gene selection from microarray data, University of South Florida, 2010.

[16] J. Canul-Reich, L.O. Hall, D. Goldgof, S.A. Eschrich, Feature selection for microarray data by AUC analysis, Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern. (2008) 768–773.

[17] C.C. Chang, C.J. Lin, LIBSVM: A Library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011).

[18] V.K. Chauhan, K. Dahiya, A. Sharma, Problem formulations and solvers in linear SVM: a review, Artif. Intell. Rev. 52 (2019) 803–855.

[19] D. Chicco, Ten quick tips for machine learning in computational biology, BioData Min. (2017).

[20] K.R. CLARKE, Non-parametric multivariate analyses of changes in community structure, Aust. J. Ecol. 18 (1993) 117–143.

[21] K.R. Clarke, R.N. Gorley, Getting started with PRIMER v7 Plymouth Routines In Multivariate Ecological Research, (2015).

[22] M.C. Collado, S. Rautava, J. Aakko, E. Isolauri, S. Salminen, Human gut colonisation may be initiated in utero by distinct microbial communities in the placenta and amniotic fluid, Sci. Rep. 6 (2016) 1–13.

[23] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, 2000.

[24] V. D'Argenio, Human microbiome acquisition and bioinformatic challenges in metagenomic studies, Int. J. Mol. Sci. 19 (2018).

[25] M. Dash, H. Liu, Consistency-based search in feature selection, Artif. Intell. 151 (2003) 155–176.

[26] H. Deng, Guided Random Forest in the RRF Package, (2013) 1–2.

[27] H. Deng, G. Runger, Feature selection via regularized trees, Proc. Int. Jt. Conf. Neural Networks. (2012).

[28] H. Deng, G. Runger, Gene selection with guided regularized random forest, Pattern

Recognit. 46 (2013) 3483–3489.

[29] W. Duch, K. Grabczewski, T. Winiarski, J. Biesiada, A. Kachel, Feature selection based on information theory, consistency and separability indices, ICONIP 2002 - Proc. 9th Int. Conf. Neural Inf. Process. Comput. Intell. E-Age. 4 (2002) 1951–1955.

[30] S.A. Dudani, The Distance-Weighted k-Nearest-Neighbor Rule, IEEE Trans. Syst. Man. Cybern. SMC-6 (1976) 325–327.

[31] D.A. Eschenbach, P.R. Davick, B.L. Williams, S.J. Klebanoff, K. Young-Smith, C.M. Critchlow, K.K. Holmes, Prevalence of hydrogen peroxide-producing Lactobacillus species in normal women and women with bacterial vaginosis, J. Clin. Microbiol. 27 (1989) 251–256.

[32] R. Fátima Medina Merino, C. Ismelda Ñique Chacón, Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python, (2017) 165–190.

[33] J. Ferlay, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, F. Bray, Global Cancer Observatory: Cancer Today, Lyon, Fr. Int. Agency Res. Cancer. (2018).

[34] H.L. Gardner, C.D. Dukes, Haemophilus vaginalis vaginitis. A newly defined specific infection previously classified “nonspecific” vaginitis, Am. J. Obstet. Gynecol. 69 (1955) 962–976.

[35] S. Goswami, A. Chakrabarti, Feature Selection: A Practitioner View, Int. J. Inf. Technol. Comput. Sci. 6 (2014) 66–77.

[36] Q. Gu, Z. Li, J. Han, Generalized fisher score for feature selection, in: Proc. 27th Conf. Uncertain. Artif. Intell. UAI 2011, 2011: pp. 266–273.

[37] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.

[38] I. Guyon, S. Gunn, M. Nikravesh, L.A. Zadeh, Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing), Series Stu, Springer, 2006.

[39] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.

[40] M.A. Hall, Correlation-based feature selection for machine learning, University of Waikato, 1999.

[41] E. Halliday, S.L. McLellan, L.A. Amaral-Zettler, M.L. Sogin, R.J. Gast, Comparison of bacterial communities in sands and water at beaches with bacterial water quality violations, PLoS One. (2014).

[42] J. Han, M. Kamber, J. Pei, Data Mining : Concepts and Techniques : Concepts and Techniques (3rd Edition), 3rd., Elsevier Amsterdam, Champaign, IL., 2012.

[43] A.U. Haq, J. Li, M.H. Memon, M. Hunain Memon, J. Khan, S.M. Marium, Heart Disease Prediction System Using Model Of Machine Learning and Sequential Backward Selection Algorithm for Features Selection, in: 2019 IEEE 5th Int. Conf. Converg. Technol., IEEE, 2019: pp. 1–4.

[44] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: data mining, inference, and prediction, Second edi, Springer Science & Business Media, 2009.

[45] J. Hernández-Torruco, Descriptive and predictive models for Guillain-Barre syndrome based on clinical data using machine learning algorithms, Universidad Juarez Autonoma de Tabasco, 2016.

[46] D.W. Hilbert, J.A. Schuyler, M.E. Adelson, E. Mordechai, J.D. Sobel, S.E. Gygax, Gardnerella vaginalis population dynamics in bacterial vaginosis, Eur. J. Clin. Microbiol. Infect. Dis. 36 (2017) 1269–1278.

[47] G.B. Hill, The microbiology of bacterial vaginosis, Am. J. Obstet. Gynecol. 169 (1993) 450–454.

[48] K. Hornik, C. Buchta, A. Zeileis, Open-source machine learning: R meets Weka, Comput. Stat. 24 (2009) 225–232.

[49] S. Hunter, M. Corbett, H. Denise, M. Fraser, A. Gonzalez-Beltran, C. Hunter, P. Jones, R. Leinonen, C. McAnulla, E. Maguire, J. Maslen, A. Mitchell, G. Nuka, A. Oisel, S. Pesseat, R. Radhakrishnan, P. Rocca-Serra, M. Scheremetjew, P. Sterk, D. Vaughan, G. Cochrane, D. Field, S.-A. Sansone, EBI metagenomics--a new resource for the analysis and archiving of metagenomic data., Nucleic Acids Res. 42 (2014) D600-6.

[50] F. Jafarnejad, S. Nayeban, K. Ghazvini, Diagnostic value of Amsel's clinical criteria for diagnosis of bacterial vaginosis, Iran. J. Obstet. Gynecol. Infertil. 13 (2010) 33–38.

[51] A. Jović, K. Brkić, N. Bogunović, A review of feature selection methods with applications, 2015 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2015 - Proc. (2015) 1200–1205.

[52] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324.

[53] I. Kononenko, On Biases in Estimating Multi-Valued Attributes, Proc. 14th Int. Jt. Conf. Artif. Intell. (1995) 1034–1040.

[54] A. Koul, C. Becchio, A. Cavallo, PredPsych: A toolbox for predictive machine learning-based approach in experimental psychology research, Behav. Res. Methods. 50 (2018) 1657–1672.

[55] M. Kuhn, Building predictive models in R using the caret package, J. Stat. Softw. 28 (2008) 1–26.

[56] K. Kumar, G. Kumar, Y. Kumar, Feature Selection Approach for Intrusion Detection System, Int. J. Adv. Trends Comput. Sci. Eng. 2 (2013) 47–53.

[57] M.B. Kursa, W.R. Rudnicki, Feature selection with the boruta package, J. Stat. Softw. 36 (2010) 1–13.

[58] S.M.R. Lannon, K.M. Adams Waldorf, T. Fiedler, R.P. Kapur, K. Agnew, L. Rajagopal, M.G. Gravett, D.N. Fredricks, Parallel detection of lactobacillus and bacterial vaginosis-associated bacterial DNA in the chorioamnion and vagina of pregnant women at term, J. Matern. Neonatal Med. 32 (2019) 2702–2710.

[59] M.Y. Lee, C.S. Yang, Entropy-based feature extraction and decision tree induction for breast cancer diagnosis with standardized thermograph images, Comput. Methods Programs Biomed. 100 (2010) 269–282.

[60] H. Liang, B.Y. Tsui, H. Ni, C.C.S. Valentim, S.L. Baxter, G. Liu, W. Cai, D.S. Kermany, X. Sun, J. Chen, L. He, J. Zhu, P. Tian, H. Shao, L. Zheng, R. Hou, S. Hewett, G. Li, P. Liang, X. Zang, Z. Zhang, L. Pan, H. Cai, R. Ling, S. Li, Y. Cui, S. Tang, H. Ye, X. Huang,

W. He, W. Liang, Q. Zhang, J. Jiang, W. Yu, J. Gao, W. Ou, Y. Deng, Q. Hou, B. Wang, C. Yao, Y. Liang, S. Zhang, Y. Duan, R. Zhang, S. Gibson, C.L. Zhang, O. Li, E.D. Zhang, G. Karin, N. Nguyen, X. Wu, C. Wen, J. Xu, W. Xu, B. Wang, W. Wang, J. Li, B. Pizzato, C. Bao, D. Xiang, W. He, S. He, Y. Zhou, W. Haw, M. Goldbaum, A. Tremoulet, C.N. Hsu, H. Carter, L. Zhu, K. Zhang, H. Xia, Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence, Nat. Med. (2019).

[61] X. Lin, C. Li, W. Ren, X. Luo, Y. Qi, A new feature selection method based on symmetrical uncertainty and interaction gain, Comput. Biol. Chem. 83 (2019) 107149.

[62] H. Liu, E.R. Dougherty, J.G. Dy, K. Torkkola, E. Tuv, H. Peng, C. Ding, F. Long, M. Berens, L. Parsons, L. Yu, Z. Zhao, G. Forman, Evolving feature selection, IEEE Intell. Syst. 20 (2005) 64–76.

[63] H. Liu, H. Motoda, Computational Methods of Feature Selection, Chapman & Hall/CRC Press, 2007.

[64] K. Lu, R.P. Abo, K.A. Schlieper, M.E. Graffam, S. Levine, J.S. Wishnok, J.A. Swenberg, S.R. Tannenbaum, J.G. Fox, Arsenic exposure perturbs the gut microbiome and its metabolic profile in mice: An integrated metagenomics and metabolomics analysis, Environ. Health Perspect. 122 (2014) 284–291.

[65] N.M. Luscombe, D. Greenbaum, M. Gerstein, A Proposed Definition and Overview of the Field, Methods Inf. Med. 40 (2001) 346–358.

[66] R. Maldonado, Sebastián; Weber, Modelos de Selección de Atributos para Support Vector Machines, Rev. Ing. Sist. (2012) 49–70.

[67] A. Marcano-Cedeno, J. Quintanilla-Dominguez, M.G. Cortina-Januchs, D. Andina, Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network, in: IECON 2010 - 36th Annu. Conf. IEEE Ind. Electron. Soc., IEEE, 2010: pp. 2845–2850.

[68] I.H. McHardy, M. Goudarzi, M. Tong, P.M. Ruegger, E. Schwager, J.R. Weger, T.G. Graeber, J.L. Sonnenburg, S. Horvath, C. Huttenhower, D.P.B. McGovern, A.J. Fornace, J. Borneman, J. Braun, Integrative analysis of the microbiome and metabolome of the

human intestinal mucosal surface reveals exquisite inter-relationships, Microbiome. 1 (2013).

[69] E. (Centro N. de I.O. Molina, Sebbm divulgación la ciencia al alcance de la mano, SEBBM. (2018) 0–1.

[70] A.B. Onderdonk, M.L. Delaney, R.N. Fichorova, The human microbiome during bacterial vaginosis, Clin. Microbiol. Rev. 29 (2016) 223–238.

[71] J.H. Orallo, M.J. Quintana, C. Ramírez, Introducción a la Minería de Datos, Pearson Educación, 2004.

[72] J.F. Perez-Gomez, J. Canul-Reich, E. Hernandez-De la Cruz, Combinación de Rankings como Método para la Identificación de Biomarcadores de Vaginosis Bacteriana, Res. Comput. Sci. (2020).

[73] B. Pes, Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains, Neural Comput. Appl. 32 (2020) 5951–5973.

[74] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Selections from Numerical Recipes in Fortran, (1993).

[75] P. Pudil, J. Novovieova, J. Kittler, Floating search methods in feature selection, Pattern Recognit. Lett. 15 (1994) 1119–1125.

[76] J. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.

[77] J.R. Quinlan, C4.5: Programs for Machine Learning., Morgan Kauf- mann Publishers, 2014.

[78] J. Ravel, P. Gajer, Z. Abdo, G.M. Schneider, S.S.K. Koenig, S.L. McCulle, S. Karlebach, R. Gorle, J. Russell, C.O. Tacket, R.M. Brotman, C.C. Davis, K. Ault, L. Peralta, L.J. Forney, Vaginal microbiome of reproductive-age women, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 4680–4687.

[79] Real Academia Española, Diccionario de la lengua española, (2020).

[80] J. Rissanen, Modelling by the shortest data description, Automatica. 14 (1978) 465–471.

[81] M. Robnik-Sikonja, F. Kononenko, Theoretical and Empirical Analysis of ReliefF and RReliefF, Mach. Learn. 53 (2003) 23–69.

[82] M. Robnik-Sikonja, P. Savicky, Package ‘ CORElearn ,’ (2021).

[83] P. Romanski, Package “FSelector,” Http://Cran.r-Project.Org/Web/Packages/FSelector/FSelector.Pdf. (2013).

[84] P. Romanski, Package “FSelector,” Http://Cran.r-Project.Org/Web/Packages/FSelector/FSelector.Pdf. (2013).

[85] D. La Rosa Hernández, E.J. Gómez Cabeza, N. Sánchez Castañeda, Intestinal microbiota in the development of the neonate’s immune system, Rev. Cubana Pediatr. 86 (2014) 502–513.

[86] R.P. Rubido, Una revisión a algoritmos de selección de atributos que tratan la redundancia en datos microarreglos, Rev. Cuba. Ciencias Informáticas. 7 (2013) 16–30.

[87] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics. 23 (2007) 2507–2517.

[88] E.K. Sanchez-garcia, A. Contreras-paredes, E. Martinez-abundis, D. Garcia-chan, E. De La Cruz-Hernandez, Molecular Epidemiology of Bacterial Vaginosis and Its Association with Sexually Transmitted Pathogens in Healthy Women, J. Med. Microbiol. Mol. 68 (2019).

[89] I.N.M. Shaharanee, F. Hadzic, Feature Selection for Data and Pattern Recognition, Stud. Comput. Intell. 584 (2015) 199–228.

[90] L. Sheugh, S.H. Alizadeh, A note on pearson correlation coefficient as a metric of similarity in recommender system, 2015 AI Robot. IRANOPEN 2015 - 5th Conf. Artif. Intell. Robot. (2015).

[91] K. Shin, X.M. Xu, Consistency-Based Feature Selection, in: Lect. Notes Comput. Sci., Springer Berlin Heidelberg, 2009: pp. 342–350.

[92] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent, J. Stat. Softw. 39 (2011) 139–148.

[93] J.D. Sobel, Bacterial Vaginosis, Annu. Rev. Med. 51 (2000) 349–356.

[94] S. Srinivasan, D.N. Fredricks, The Human Vaginal Bacterial Biota and Bacterial Vaginosis, Interdiscip. Perspect. Infect. Dis. 2008 (2008) 1–22.

[95] S. Srinivasan, N.G. Hoffman, M.T. Morgan, F.A. Matsen, T.L. Fiedler, R.W. Hall, F.J. Ross, C.O. McCoy, R. Bumgarner, J.M. Marrazzo, D.N. Fredricks, Bacterial communities in women with bacterial vaginosis: High resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria, PLoS One. 7 (2012).

[96] K. Strimbu, J.A. Tavel, What are biomarkers?, Curr. Opin. HIV AIDS. 5 (2010) 463–466.

[97] R. Tibshirani, A proposal for variable selection in the cox model, Stat. Med. 0 (1994) 1–12.

[98] L. Tipton, K.T. Cuenco, L. Huang, R.M. Greenblatt, E. Kleerup, F. Sciurba, S.R. Duncan, M.P. Donahoe, A. Morris, E. Ghedin, Measuring associations between the microbiota and repeated measures of continuous clinical variables using a lasso-penalized generalized linear mixed model, BioData Min. 11 (2018) 1–20.

[99] L. Torgo, Data Mining with R, learning with case studies, Chapman and Hall/CRC., 2010.

[100] R.J. Urbanowicz, M. Meeker, W. La Cava, R.S. Olson, J.H. Moore, Relief-based feature selection: Introduction and review, J. Biomed. Inform. 85 (2018) 189–203.

[101] M.A. Valizade Hasanloei, R. Sheikhpour, M.A. Sarram, E. Sheikhpour, H. Sharifi, A combined Fisher and Laplacian score for feature selection in QSAR based drug design using compounds with known and unknown activities, J. Comput. Aided. Mol. Des. 32 (2018) 375–384.

[102] H. Wang, B. Zheng, S.W. Yoon, H.S. Ko, A support vector machine-based ensemble algorithm for breast cancer diagnosis, Eur. J. Oper. Res. 267 (2018) 687–699.

[103] Q. Wang, G.M. Garrity, J.M. Tiedje, J.R. Cole, Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy, Appl. Environ. Microbiol. 73 (2007) 5261–5267.

[104] D.I. Warton, S.T. Wright, Y. Wang, Distance-based multivariate analyses confound location and dispersion effects, Methods Ecol. Evol. 3 (2012) 89–101.

[105] C.M. Wheeler, Natural History of Human Papillomavirus Infections, Cytologic and Histologic Abnormalities, and Cancer, Obstet. Gynecol. Clin. North Am. 35 (2008) 519–536.

[106] WHO, International Programme on Chemical Safety, Biomarkers in risk assessment:

validity and validation, Biomarkers Risk Assess. Validity Valid. (2001).

[107] J.C.F. de Winter, S.D. Gosling, J. Potter, Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data., Psychol. Methods. 21 (2016) 273–290.

[108] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, (2005).

[109] I.H. Witten, E. Frank, J. Geller, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Elsevier, 2002.

[110] A. Wosiak, D. Zakrzewska, Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis, Complexity. 2018 (2018).