Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Interacciones del microbioma intestinal humano en diferentes comunidades europeas: un enfoque complementario desde el aprendizaje automático

Tesis presentada para optar al título de Magister de la Universidad de Buenos Aires en Explotación de Datos y el Descubrimiento de Conocimiento

Dra. Valeria Laura Burgos

Directora de tesis: Dra. María Laura Fernández

Co-Director: Dr. Marcelo Risk

Tesis desarrollada en el Instituto de Medicina Traslacional e Ingeniería Biomédica (IMTIB) CONICET - Instituto Universitario Hospital Italiano - Hospital Italiano de Buenos Aires

Buenos Aires, Diciembre de 2021

Dedicatoria

Esta tesis está dedicada a la memoria del Dr. Pablo F. Argibay.

Agradecimientos

A mi directora, María Laura Fernández, por acompañarme y aconsejarme en las diversas fases por las que atravesó la realización de esta tesis.

A mi co-director, Marcelo Risk, por su constante ayuda y estímulo para crecer profesionalmente.

A Pilar Ávila, Amalia Guaymás y a mis compañeros de cursada que hicieron que la maestría haya sido una gran experiencia.

A mi hija Sophia.

A mis amigos Guido Guzmán y Nicolás Quiroz.

Resumen

El tracto gastrointestinal humano está colonizado por una abundancia de comunidades de bacterias, virus, hongos y arqueas, conviviendo en equilibrio entre sí y con el hospedador humano. Son millones de microorganismos que forman la denominada microbiota intestinal, con roles muy importantes en el bienestar, mantenimiento de la salud y aparición de varias enfermedades que van desde el síndrome de colon irritable, cáncer hasta depresión.

En la presente Tesis se propone un enfoque de análisis desde *data mining* sobre un conjunto de datos de microbioma intestinal provenientes de individuos europeos sanos, utilizando algoritmos de *machine learning* que permitan identificar relaciones con relevancia biomédica entre las variables demográficas de los sujetos de estudio y algún patrón de abundancia de bacterias.

El estudio se realizó sobre las varias especies de bacterias pertenecientes a los dos grandes grupos (o *phyla*) de mayor abundancia en la microbiota intestinal humana, tal como *Bacteroidetes* y *Firmicutes*, teniendo en cuenta además información sobre el el grado de obesidad o delgadez de los individuos. Ambos subconjuntos de abundancias de bacterias junto a las variables de metadata fueron sujetos a un análisis exploratorio, usando posteriormente análisis de componentes principales y algoritmos de *clustering* con el fin de caracterizar la presencia de *clusters* naturales en los datos, además de evaluar la variabilidad de los mismos.

Se usó también el algoritmo de mapas auto-organizados cuya visualización de los resultados representa una valiosa herramienta de análisis ya que permite integrar información multivariada a través de diferentes planos componentes.

Por último, para estudiar la diversidad de las poblaciones de bacterias en estudio y su asociación con datos demográficos de los individuos, se usaron dos métodos de ensamble como *random forest* y *gradient boosting machine* para modelar la importancia de las variables sobre la diversidad biológica de las comunidades.

Los resultados indicaron que la edad y el índice de masa corporal fueron las variables más importantes en explicar la diversidad de bacterias. Curiosamente, varias especies de bacterias conocidas por estar asociados con la dieta y la obesidad, fueron identificadas como características relevantes. El análisis topológico de los mapas auto-organizados identificó ciertos grupos de nodos de características similares en los metadatos de los sujetos y grupos de bacterias.

Se concluye que los resultados representan un enfoque que podría ser considerado, en futuros estudios, como un potencial complemento en reportes de salud para ayudar a los profesionales de la salud a personalizar el tratamiento del paciente o apoyo a la toma de decisiones.

Palabras clave: microbiota intestinal, mapas auto-organizados, *random forest*, *grandient boosting machine*, medicina personalizada, bioinformática.

Abstract

The human gastrointestinal tract is colonized by an abundance of communities of bacteria, viruses, fungi and archaea, living in balance with each other and with the human host. There are millions of microorganisms that make up the so-called gut microbiota, with very important roles in the well-being, health maintenance as well as the appearance of various diseases ranging from irritable bowel syndrome, cancer to depression.

This thesis proposes a data mining approach to analyse a set of gut microbiota data from healthy individuals, using machine learning algorithms to identify biomedically relevant relationships between demographic and biomedical variables of the subjects and patterns of abundance of bacteria.

The study was carried out focusing on the two most abundant human gut microbiota groups (or phyla), such as *Bacteroidetes* and *Firmicutes*. Both subsets of bacterial abundances together with the metadata variables were subjected to an exploratory analysis, subsequently using principal component analysis and clustering algorithms in order to characterize the presence of natural clusters in the data, in addition to evaluating their variability.

The self-organizing map algorithm was also used, whose visualization power represents a valuable analysis tool since it integrates multivariate information through different component planes.

Finally, to evaluate the relevance of the variables on the biological diversity of the microbial communities, two ensemble-based methods such as random forest and gradient boosting machine were used.

Results showed that age and body mass index were the most important features at explaining bacteria diversity. Interestingly, several species of bacteria known to be associated to diet and obesity were identified as relevant features as well. In the topological analysis of self-organizing maps, we identified certain groups of nodes with similarities in subject metadata and gut bacteria.

It is concluded that the results represent an approach that could be considered, in future studies, as a potential complement in health reports so as to help health professionals personalize patient treatment or support decision-making.

Keywords: gut microbiota, self-organizing maps, random forest, gradient boosting machine, personalized medicine, bioinformatics.

Lista de abreviaturas

ADN Ácido Desoxiribonucleico

ARN Ácido Ribonubleico

BMI *Body Mass Index* – Índice de Masa Corporal

GBDT *Gradient Boosting Decision Tree*

GI Gastrointestinal

HGP *Human Genome Project* – Proyecto Genoma Humano

MAE *Mean Absolute Error* – Error Absoluto Medio

PC *Principal Component* – Componente Principal

PCA *Principal Component Analysis* – Análisis de Componentes Principales

ML *Machine Learning* – Aprendizaje Automático

NA *Not Available* – Valor faltante

NGS *Next Generation Sequencing*

RF *Random Forest*

SOM *Self-Organizing Maps* – Mapas Auto-Organizados

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Los seres vivos están clasificados dentro de tres grandes dominios: Bacterias, Arqueas y Eucariotas, tal cual indica el modelo conocido como árbol de la vida, que explora la evolución y las relaciones de todos los organismos (Figura 1.1). Una rama del árbol representa a los eucariotas, que incluye animales, plantas, hongos, algas y protozoos. La segunda rama describe a las bacterias; microorganismos unicelulares (algunas de los cuales serán detallados más adelante). Y la tercer rama abarca a arqueas, microorganismos presentes en ambientes considerados extremos para la vida (en términos de temperatura, pH, salinidad y anaerobiosis) y no extremos, como océanos, sedimentos de agua dulce y suelos [Chaban et al., 2006].

Desde organismos simples compuestos por una única célula (como arqueas, bacterias, protozoos, algunas algas y hongos) hasta los complejos organismos multicelulares (tal como los humanos), las instrucciones necesarias para la creación de biomoléculas y mantenimiento de funciones celulares durante la vida de un organismo están codificadas en la molécula de ácido desoxiribonucleico (ADN). Esta molécula está formada por un par de cadenas largas pareadas, conformada por cuatro tipos de monómeros (Adenina [A], Timina [T], Citosina [C] y Guanina [G]) unidos en una secuencia lineal. El ADN contiene las instrucciones biológicas para producir, por ejemplo, proteínas (un tipo de biomolécula con diversos roles en la célula, tal como formar estructuras biológicas). También se incluyen instrucciones para mecanismos de reproducción, funciones especializadas, interacción con otras células y respuestas a señales del entorno,

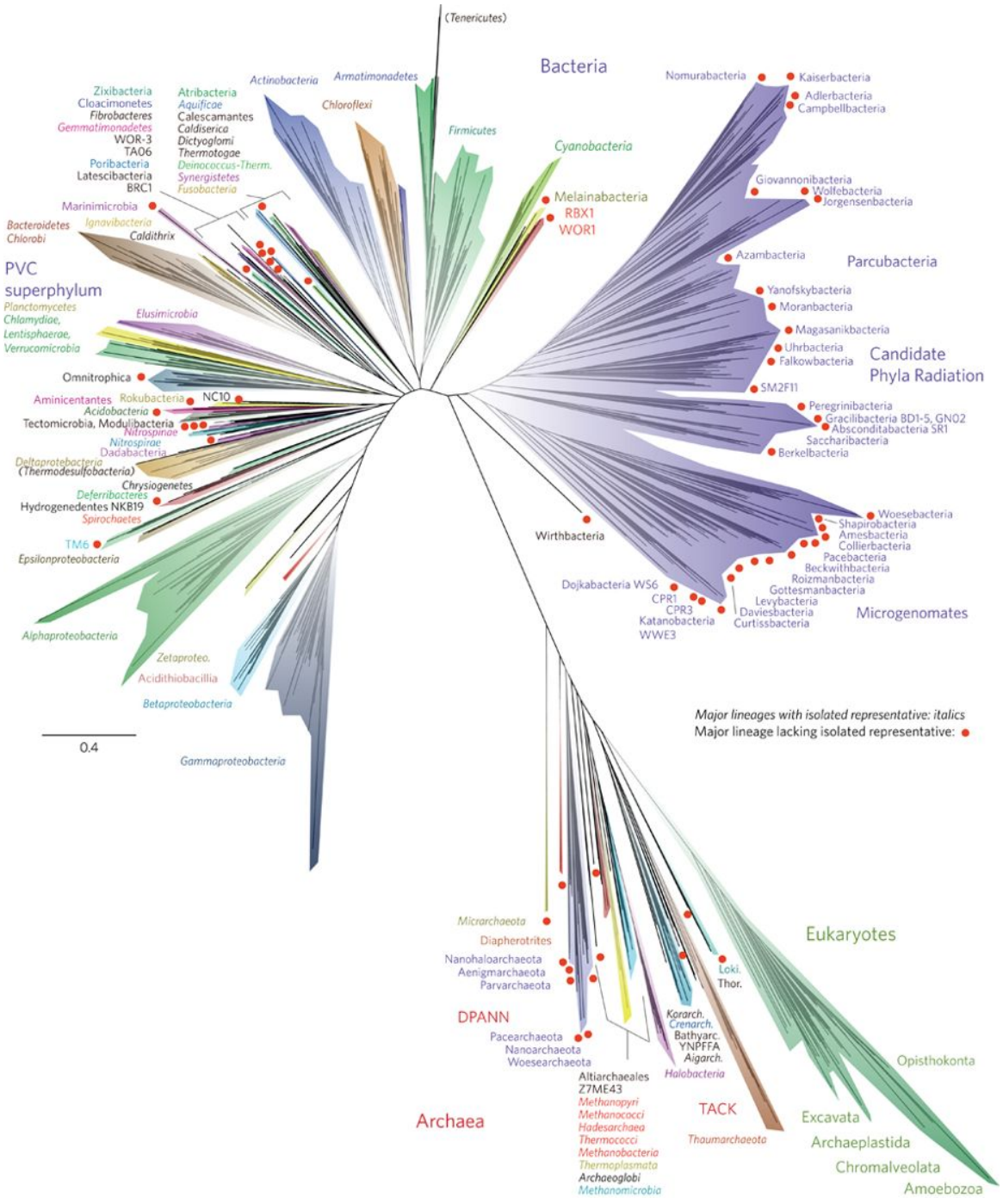

Figura 1.1: Esquema actual del árbol de la vida, comprendiendo la diversidad total representada por genomas secuenciados. El árbol incluye 92 nombres de divisiones (*phyla*) de Bacterias, 26 *phyla* de Arquea y los 5 supergrupos de Eucariotas. Adaptado de [Hug et al., 2016].

etc. En definitiva, las instrucciones para mantener la vida y hacer única a cada especie.

Estas instrucciones en forma de combinaciones de los cuatro monómeros A, T, C y G (conocidos como nucleótidos) están almacenadas en forma de **gen**, definido como una unidad funcional de ADN que codifica moléculas cuyas funciones dirigen todas las actividades de la vida [Alberts et al., 2002]. Esta codificación de la información

genética es similar a la brindada por la secuencia de unos y ceros en archivos de computadora. Los genes funcionan como elementos contenedores de información que determinan las características de una especie en su totalidad y de los individuos dentro de la misma.

Teniendo en cuenta que esta Tesis estudia la abundancia de bacterias intestinales y su interacción con el hospedador humano, es necesario describir muy brevemente algunas diferencias entre ambos tipos de organismos desde el punto de vista de la organización celular de la molécula de ADN.

En el interior celular de las bacterias, el ADN se encuentra libre, sin un compartimento nuclear definido que lo separe del resto de los componentes intracelulares. Y existe como molécula desnuda, lo que quiere decir que no está acompañado en su estructura por ningún complejo proteico u otras macromoléculas. Estas características son típicas de organismos clasificados como Procariotas, a los cuales pertenecen Bacterias y Arqueas (Figura 1.2).

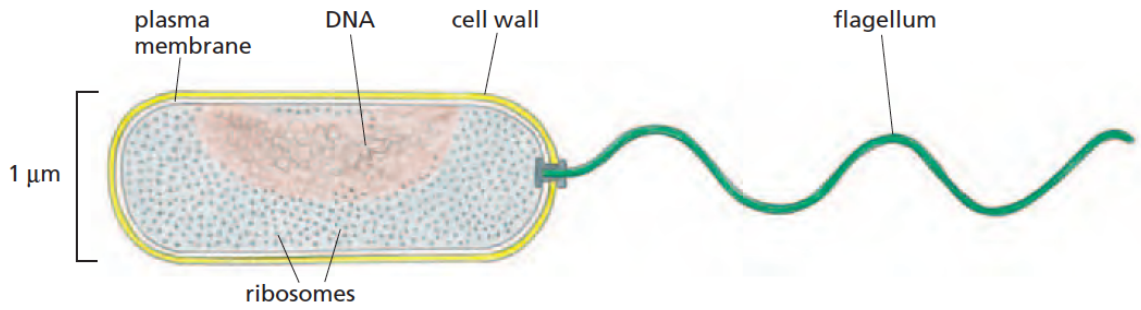

Figura 1.2: Esquema de una célula procariota. En este ejemplo, la estructura de la bacteria *Vibrio cholerae* mostrando su organización interna simple. Como muchas otras especies, *V. cholerae* tiene un flagelo que rota como propulsor de movimiento de la célula. Adaptado de [Alberts et al., 2002].

En comparación, en el interior de una célula eucariota (por ejemplo, la humana) la molécula de ADN sí está en un compartimento definido llamado núcleo, compuesto por una membrana nuclear que permite la comunicación y tránsito de moléculas entre el espacio intracelular y el interior del núcleo (Figura 1.3). Además, el ADN está acompañado en su estructura por un complejo proteico formando cromosomas (Figura 1.4). Estas son características de organismos clasificados como Eucariotas (a los cuales pertenece el ser humano).

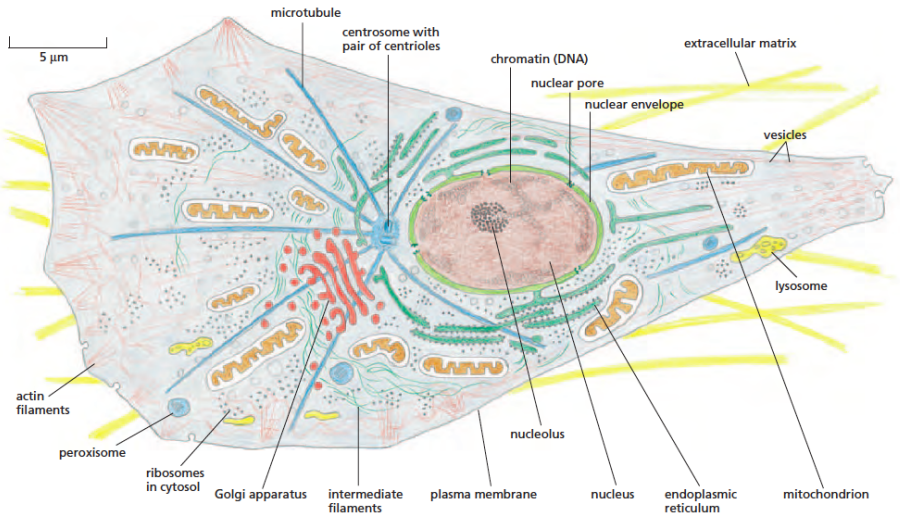

Figura 1.3: Esquema de una célula eucariota. El esquema corresponde a una célula animal, pero casi todos los componentes se encuentran también en plantas y hongos, además de eucariotas unicelulares como levaduras y protozoos. Adaptado de [Alberts et al., 2002].

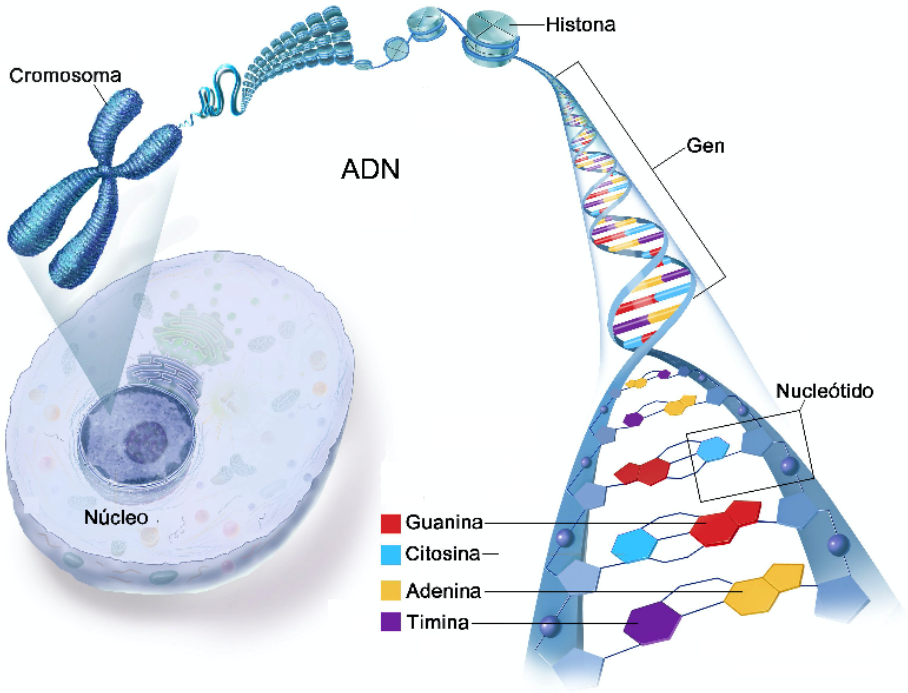

Figura 1.4: Estructura del ADN en una célula eucariota. El ADN se encuentra en el interior del núcleo de una célula, donde forma los cromosomas. Los cromosomas contienen proteínas llamadas histonas sobre las cuales se enrolla el ADN. Adaptado de www.cancer.gov.

En las diferencias mencionadas a nivel de organización del ADN subyacen a su vez diferentes y complejos mecanismos de regulación de expresión de genes en un producto funcional. Sin embargo, en todas las células de organismos vivos (tanto eucariotas como procariotas) rige un principio (en gran parte fundamental, aunque

existen excepciones) denominado *dogma central de la biologia molecular*. El mismo se refiere al flujo principal de la información genética que va desde el ADN (que contiene la información necesaria para la producción), al ácido ribonucleico (ARN, que actúa como un intermediario) y a las proteínas. Cuando la célula necesita una proteína determinada, la secuencia de nucleótidos del gen (ubicado en el ADN) es copiada primero en una molécula de ARN (proceso denominado transcripción). Estas copias de ARN son usadas directamente como moldes para dirigir la síntesis de la proteína en cuestión (proceso conocido como traducción). Por lo tanto, el flujo de información genética en las células es desde el ADN a ARN a proteína (Figura 1.5). En todas las células, la expresión de genes individuales es un proceso altamente regulado: una célula no produce el repertorio total de posibles proteínas a toda velocidad todo el tiempo sino que ajusta la tasa de transcripción y traducción de diferentes genes de manera independiente, de acuerdo a la necesidad [Alberts et al., 2002].

Esta descripción muy general sobre ciertos conceptos fundamentales de biología molecular es necesaria para brindar un contexto a términos como 'gen', 'expresión génica' y demás, dado que los datos de microbiota intestinal analizados en esta Tesis fueron obtenidos mediante una técnica de biología molecular mencionada en secciones posteriores.

1.1. Microbioma

Una gran diversidad de microorganismos abundan en las superficies e interior del cuerpo humano, desde la piel, cavidades orales y genitales, tracto respiratorio y el sistema gastrointestinal. Los microorganismos incluyen a bacterias, arqueas, virus y hongos, los que en conjunto constituyen la **microbiota** [Lloyd-Price et al., 2016]. Las comunidades de microorganismos se distribuyen en diferentes composiciones dependiendo del sitio en el hospedador (Figura 1.6).

El creciente entendimiento de la microbiota humana ha significado un cambio de paradigma en las ciencias biomédicas en los últimos años, dado que los microorganismos eran considerados principalmente como patógenos [Sansonetti, 2011]. De esta

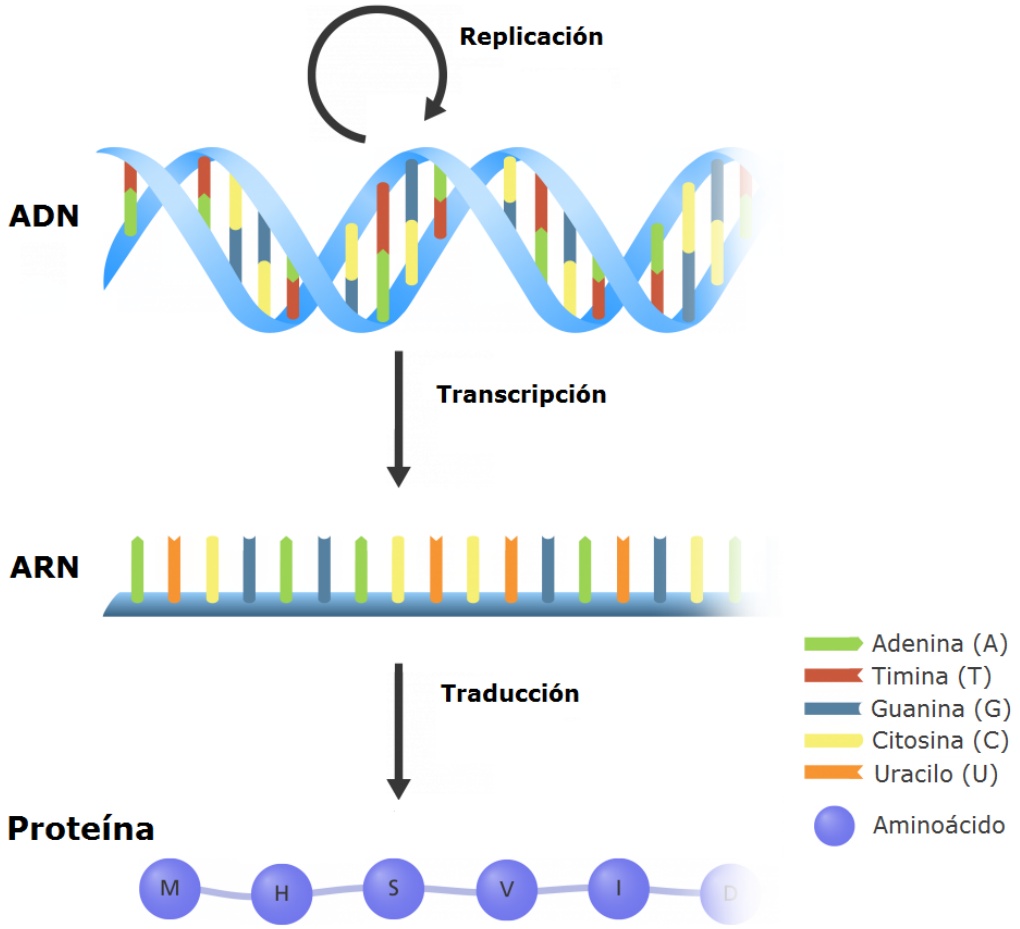

Figura 1.5: Relación entre la información genética contenida en el ADN y las proteínas.

manera, se tiene cada vez más conocimiento de las complejas y dinámicas interacciones de convivencia entre la microbiota y el hospedador, es decir, relaciones biológicas simbióticas. La simbiosis se refiere a la relación cercana y mutuamente beneficiosa entre dos organismos diferentes y abarca al comensalismo, mutualismo y parasitismo. Las bacterias tienen una larga historia de simbiosis: en el comensalismo, la relación entre ambos organismos es neutra ya que ninguna de las partes saca provecho de la otra ni se provocan perjuicio mutuo alguno. En el mutualismo, ambos organismos sacan provecho de la relación. Y en el parasitismo, un organismo se aprovecha del otro mientras lo daña [European Society Neurogastroenterology and Motility, 2020].

Existe una variedad de interacciones simbióticas entre el humano y su microbiota en el contexto del mantenimiento de la homeostasis. Por ejemplo, en el ecosistema vaginal de mujeres en edad reproductiva, las bacterias pertenecientes principalmente al

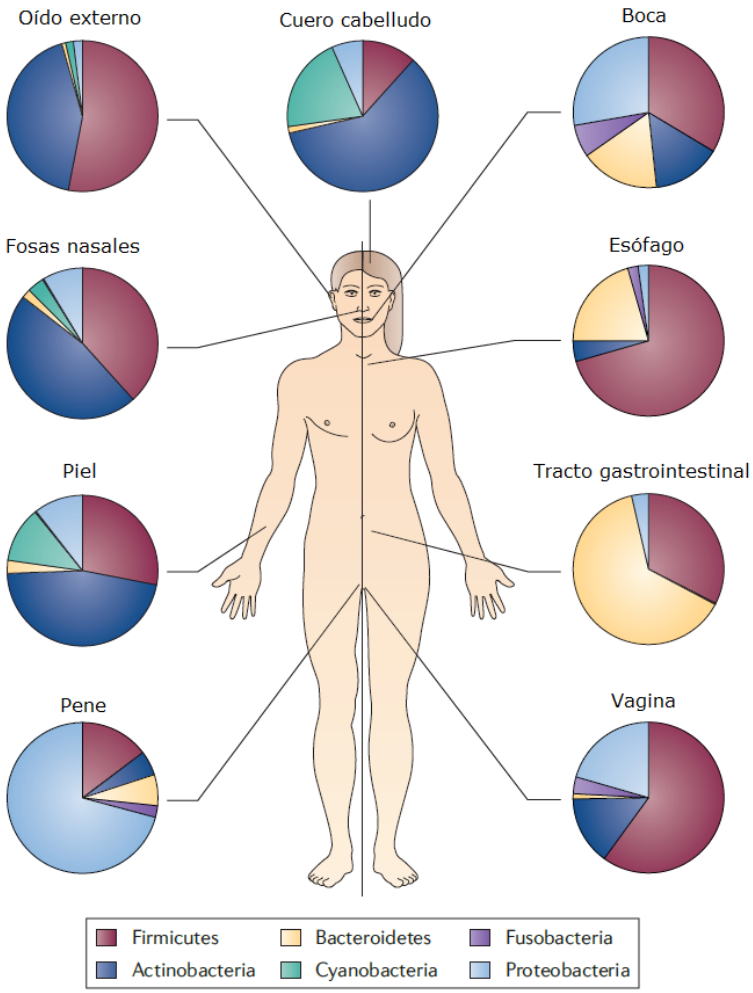

Figura 1.6: Composición de comunidades de microorganismos en diferentes partes del cuerpo en un individuo sano. Se indican las abundancias relativas de los seis grupos dominantes de bacterias. Adaptado de [Spor et al., 2011].

género *Lactobacillus* parecen tener un papel fundamental en el sistema de defensa del nicho: se hipotetiza que mediante la producción de ácido láctico, producción de compuestos bacteriostáticos y bactericidas o por exclusión competitiva se logra disminuír el pH, lo cual actúa como una protección contra la colonización excesiva por microorganismos patógenos, tal como la levadura *Candida albicans* [Ravel et al., 2011], [Tortora et al., 2007]. La composición de la microbiota en cualquier nicho del cuerpo puede ser perturbada ante la presencia de factores externos, provocando una disrupción en las relaciones simbióticas entre el hospedador y los microorganismos asociados. Este desequilibrio persistente, conocido como **disbiosis**, está asociado a enfermedades [Floch et al., 2016].

La presencia de diversas comunidades de bacterias en el cuerpo humano representa

una importante expansión funcional del genoma hospedador, es decir, un aporte de características que los humanos no necesitaron evolucionar por su cuenta (revisado en [Gill et al., 2006]). Por ejemplo, el genoma bacteriano aporta varios tipos de enzimas que participan en la digestión de ciertos compuestos que no pueden ser digeridos en el estómago, como componentes vegetales o ciertos azúcares [Larsbrink et al., 2014]. De esta manera, existe una contribución relevante al metabolismo y fisiología del hospedador por parte de algunas bacterias comensales. El genoma del total de los microorganismos que habitan los diferentes nichos de un organismo vertebrado es lo que se conoce como **microbioma**.

La microbiota es un ecosistema vivo y como tal, las tasas de crecimiento y supervivencia de sus poblaciones componentes pueden fluctuar. Se considera que la microbiota presenta dos características muy importantes: plasticidad y resiliencia [Ruggles et al., 2018, Liu et al., 2019]. La resiliencia se pone en evidencia cuando nos exponemos a agentes estresores temporales, tales como cambios en la dieta o el consumo de antibióticos: la microbiota responde adaptándose y recuperando el estado basal una vez que el cesa el estresor [Greenhalgh et al., 2016]. Sin embargo, también presenta la característica de flexibilidad, porque en su constitución influyen varios factores, tales como la genética del hospedador, el entorno, el estado nutricional, edad y estilo de vida, entre otros (Figura 1.7).

1.1.1. Microbioma intestinal

En humanos, el tracto gastrointestinal (GI) comienza en la cavidad oral y continúa a través del estómago e intestinos, finalizando en el ano.

El complejo consorcio de millones de microorganismos que habitan el tracto GI se ha convertido en un tópico de gran interés debido a su importante influencia en estados de fisiológicos o patológicos.

A lo largo del tracto GI se diferencian nichos fisiológicos asociados a una densidad y composición de microorganismos característicos. Esta estratificación espacial heterogénea se distribuye desde la boca hasta el recto y también a nivel transversal (desde

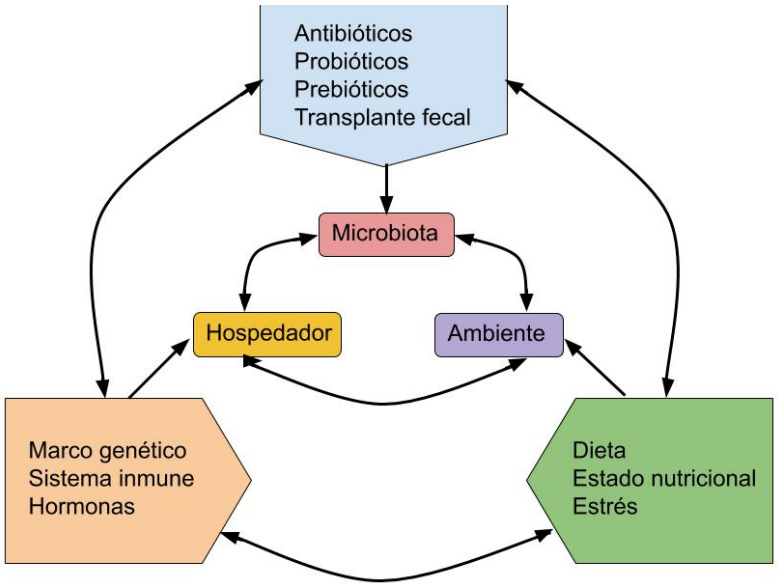

Figura 1.7: Esquema de la complejidad de interacciones entre el hospedador, la microbiota y el ambiente. Adaptado de [Eloe-Fadrosh and Rasko, 2013].

la mucosa hacia el lumen, Figura 1.8). Algunos de los factores que influyen en la heterogeneidad de distribución espacial de la microbiota del tracto GI son el pH, nivel de oxígeno, gradientes químicos y de nutrientes, péptidos antimicrobiales y características físicas del intestino, entre otros [Donaldson et al., 2016, Tropini et al., 2017].

Diversos trabajos en modelos animales y humanos han destacado la asociación de la disbiosis intestinal con una lista de enfermedades crónicas que incluyen obesidad, síndrome de colon irritable, diabetes, cáncer, y neurológicas como Parkinson, entre otras [Bäckhed et al., 2004, Matsuoka and Kanai, 2015, Dutta et al., 2019].

Mientras que muchas bacterias están asociadas con enfermedades, la microbiota intestinal tiene un participación esencial en diversas funciones beneficiosas para la salud, tal como la digestión, síntesis de vitaminas y aminoácidos esenciales, absorción de calcio, magnesio y hierro, fermentación de componentes no digeribles y y protección del intestino ante patógenos, entre otros [Egert et al., 2006, Bäckhed et al., 2005].

Se cree que durante la gestación, el feto humano se desarrolla en un ambiente libre de bacterias y al momento del parto, el recién nacido es expuesto a un amplio espectro de microorganismos que empiezan a colonizar de una manera altamente sincroniza-

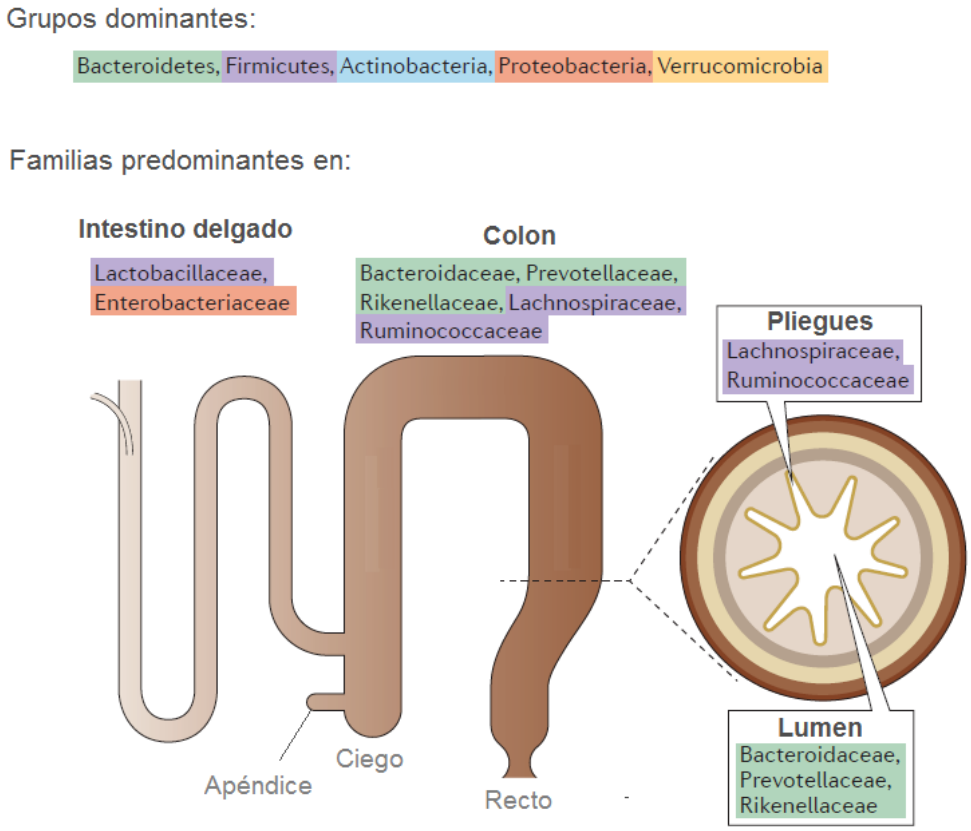

Figura 1.8: Hábitats de la microbiota en la porción baja del tracto GI. Adaptado de [Donaldson et al., 2016]

da el intestino neonatal, contribuyendo así a las funciones de protección, inmunes y metabólicas. El pasaje a través del canal de parto, el contacto con la microbiota vaginal, fecal y de piel de la madre son factores que determinan la adquisición y establecimiento de la microbiota inicial en los recién nacidos. En este sentido, existe evidencia de que los nacidos mediante parto vaginal adquieren comunidades bacterianas similares a la microbiota vaginal de su madre, con predominio de especies de *Lactobacillus*, *Prevotella* o *Sneathia spp*. Mientras que en los nacidos por cesárea predominan comunidades bacterianas similares a las encontradas en la superficie de la piel, con dominancia de *Staphylococcus*, *Corybacterium* y *Propionibacterium spp* [Dominguez-Bello et al., 2010].

La diversidad de las poblaciones que componen la microbiota intestinal es baja en los primeros períodos post-natales y va incrementando con el transcurso del tiempo. Es probable que el aumento en diversidad y riqueza de especies sea el reflejo del aumento

de tamaño del intestino, lo cual va asociado también con una mayor tasa de interacciones con nuevas bacterias y la proliferación de nichos ecológicos [Koenig et al., 2011].

El período de la infancia humana representa para la microbiota intestinal una etapa de rápida colonización de microorganismos que está influenciada por diversos factores, como tipo de parto, alimentación con leche materna o fórmula, duración del período de lactancia, introducción de alimentos sólidos, enfermedad, uso de antibióticos, etc. (Figura 1.9)

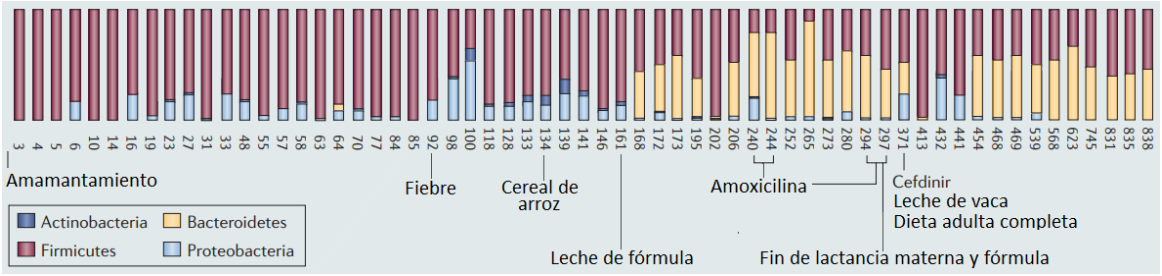

Figura 1.9: Variación de la abundancia de los principales grupos de bacterias (*Actinobacteria*, *Bacteroidetes*, *Firmicutes* y *Proteobacteria*) en el microbioma intestinal en la fase temprana de la niñez ante diversos eventos como consumo de leche materna, fórmula, antibióticos (amoxicilina y cefdinir) y alimentos procesados. Los números indican días postnacimiento. Adaptado de [Spor et al., 2011].

De esta manera, el entendimiento de los factores y mecanismos que afectan al ensamblado y mantenimiento de la microbiota intestinal permite aportar un perfil de comprensión adicional al tratamiento de enfermedades tan complejas.

1.1.2. ¿Cómo se estudia el microbioma?

La biodiversidad de la microbiota humana ha estado por muchos años subestimada debido fundamentalmente a la limitación de los métodos de aislamiento y cultivo de poblaciones de bacterias. Sin embargo, la innovación tecnológica de los últimos años ha permitido profundizar el conocimiento de las características, diversidad y funciones de la gran cantidad de microorganismos intestinales en el humano.

La investigación de la microbiota intestinal puede ser abordada mediante estudios basados en el ADN y pueden ser clasificados en dos categorías principalmente:

- Los que acceden directamente al contenido genético de comunidades enteras de microorganismos en muestras de ambiente natural (suelo, agua, intestino, etc),

área conocida como metagenómica.

- Los enfocados en uno o unos pocos genes marcadores. Esto representa una herramienta clave de estudio, dado que a través del perfil de expresión de un gen marcador, es posible caracterizar la composición y abundancia de los microorganismos presentes. Un gen marcador es un segmento de ADN del cual se conoce su ubicación en el mismo y puede ser rastreado a través de sucesivas generaciones.

En esta última categoría, el análisis del gen marcador se centra principalmente en el gen del ARN ribosomal de la subunidad pequeña, conocido como **16S rRNA**. La importancia de utilizarlo como una suerte de 'código de barras' se debe a la expresión ubicua del gen: todas las bacterias y arqueas tienen el gen 16S.

El gen 16S rRNA tiene varias regiones conservadas en su secuencia (en biología, una región conservada se refiere a una secuencias idéntica o similare entre especies mantenida por selección natural) y nueve regiones hipervariables (denominadas V1-V9) que representan suficiente diversidad de secuencia para realizar la identificación y clasificación de bacterias en diferentes grupos (Figura 1.10).

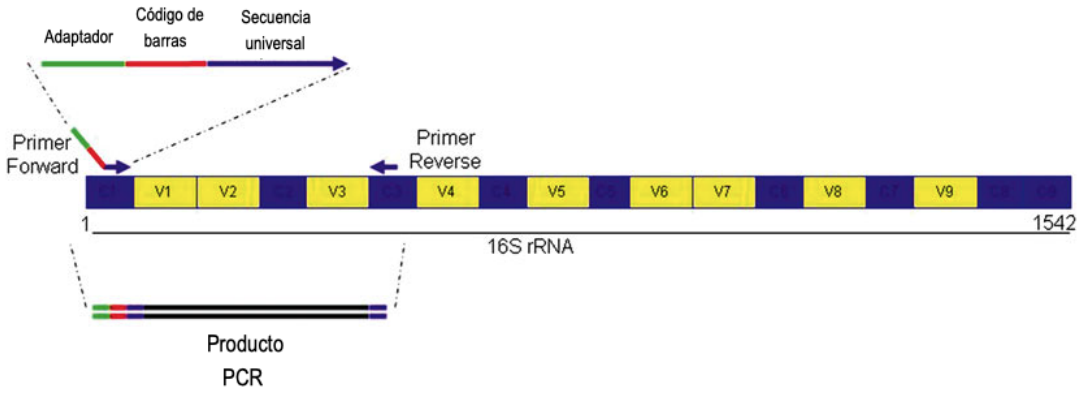

Figura 1.10: Esquema del gen 16S rRNA, indicando las regiones conservadas (en violeta) e hipervariables V1 a V9 (en amarillo). En este ejemplo, la región comprendida entre las flechas *primer forward* y *reverse* es la región *target* a amplificar y analizar (producto PCR). Las diferencias encontradas en ese fragmento amplificado permitirán determinar las diferencias de abundancias de bacterias en una muestra. Adaptado de [Del Chierico et al., 2015]

Las características mencionadas del gen permiten cuantificar las abundancias diferentes de cada grupo de bacterias en base a la variabilidad de secuencia en una determinada región hipervariable en estudio. Así, es posible identificar la composición

taxonómica de la microbiota (Figura 1.11).

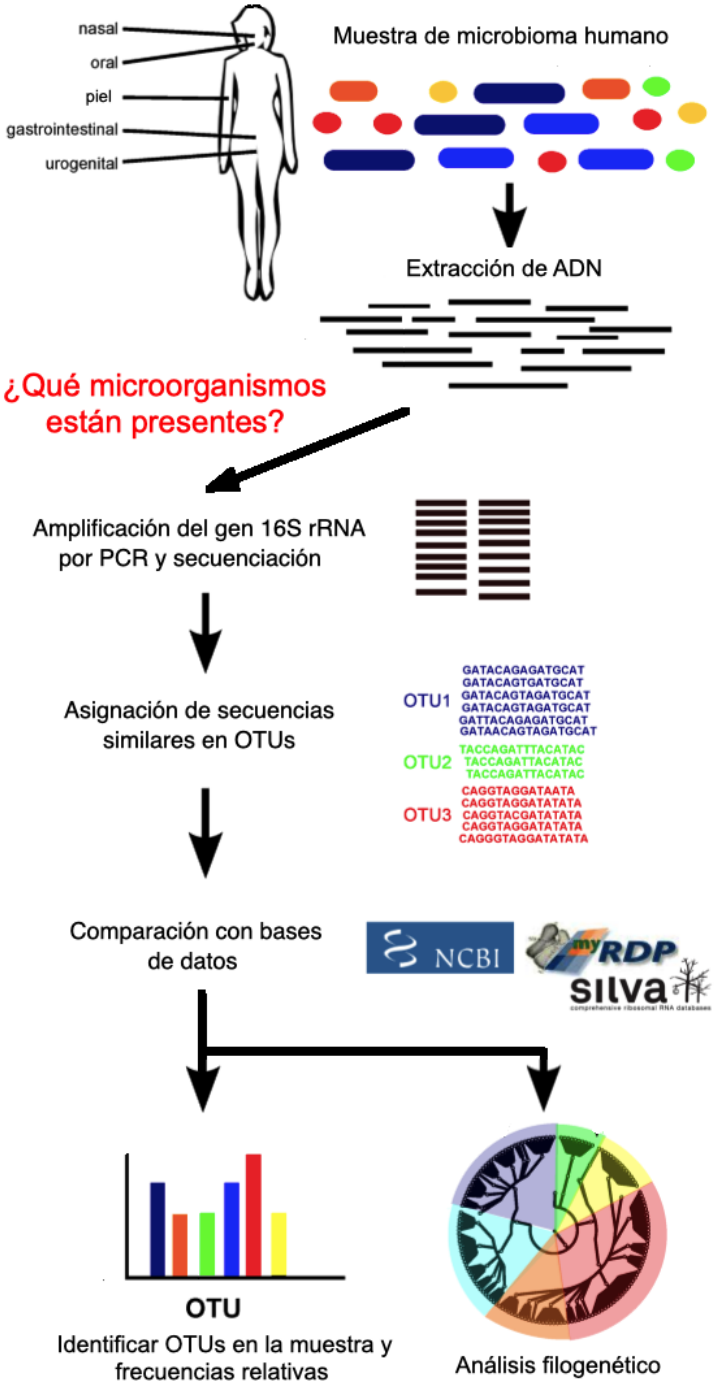

Figura 1.11: La composición de la comunidad se determina amplificando el gen 16S rRNA. Las secuencias muy similares se agrupan en unidades taxonómicas operativas (OTU), las cuales son comparadas con bases de datos de organismos reconocidos y analizadas en términos de abundancia o diversidad filogenética. Adaptado de [Morgan et al., 2013].

Las técnicas para caracterizar la microbiota intestinal dependen de métodos de biología molecular independientes del cultivo celular e incluyen:

- Tecnologías de secuenciación a gran escala, como la secuenciación de próxima generación (NGS, *Next Generation Sequencing*).
- Microarreglos o *microarrays* filogenéticos de ADN.

A continuación se describirán las características de los *microarrays* filogenéticos dado que los datos usados en esta Tesis fueron obtenidos mediante dicha técnica.

1.1.3. *Microarrays* filogenéticos

Un *microarray* emplea sondas de ADN (fragmentos cortos, de pocos nucleótidos de longitud) inmovilizadas y organizadas de manera predeterminada en forma de grilla en una placa de vidrio o sílice. Las sondas están dispuestas en una serie de puntos de alta densidad, donde cada punto corresponde a una sonda diseñada para ser complementaria a una secuencia específica de ADN proveniente de la muestra a analizar. Por ejemplo, si se quiere analizar la expresión de un gen de interés en una muestra, es posible extraer todo el conjunto de genes que son expresados en ese momento por las células (una suerte de instantánea génica). Los mismos son marcados con fluorescencia y, luego de incubar en condiciones favorables, los genes presentes que hayan hibridado con las sondas de la placa emitirán una señal de fluorescencia. La intensidad de fluorescencia emitida en cada posición en el *chip* refleja la abundancia del gen correspondiente a esa secuencia. De esta manera, es posible examinar los patrones de expresión génica global en un momento determinado (Figura 1.12).

En particular, un *microarray* filogenético contiene sondas complementarias a una determinada región hipervariable del gen 16S rRNA. Puede ser usado para determinar con exactitud las diferencias en abundancias relativas y diversidad de bacterias.

1.2. Bioinformática

En todos los sistemas biológicos, la generación de información ha sido siempre un componente natural, inherente y dinámico. La finalización del Proyecto Genoma Humano (HGP, por sus siglas en inglés) en 2003 y los posteriores resultados de se-

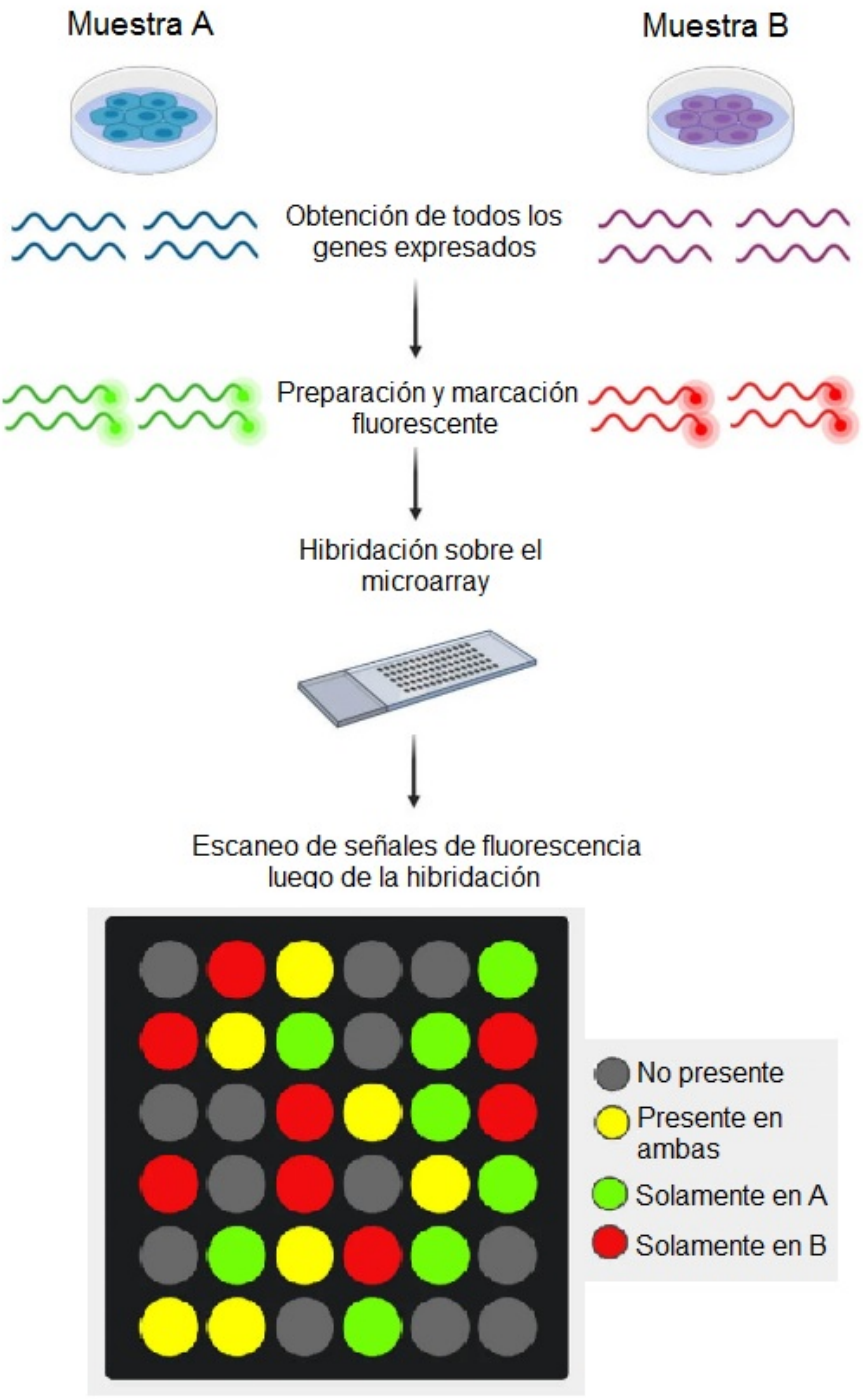

Figura 1.12: Diagrama básico de la técnica de *microarray* de ADN con dos marcas fluorescentes. Cada *spot* representa un único gen. Los *microarrays* de ADN permiten analizar la expresión de miles de genes simultáneamente.

cuenciación de genomas de otros organismos representó un cambio fundamental en las biociencias debido a la abundante disponibilidad de datos biológicos. Esta situación generó a su vez una demanda de herramientas computacionales altamente eficientes para el almacenamiento en bases de datos específicas, análisis e interpretación. La

combinación de varias disciplinas como ciencias matemáticas, estadística, ciencias de la computación, biología y medicina ha posibilitado el desarrollo de la Bioinformática (Figura 1.13).

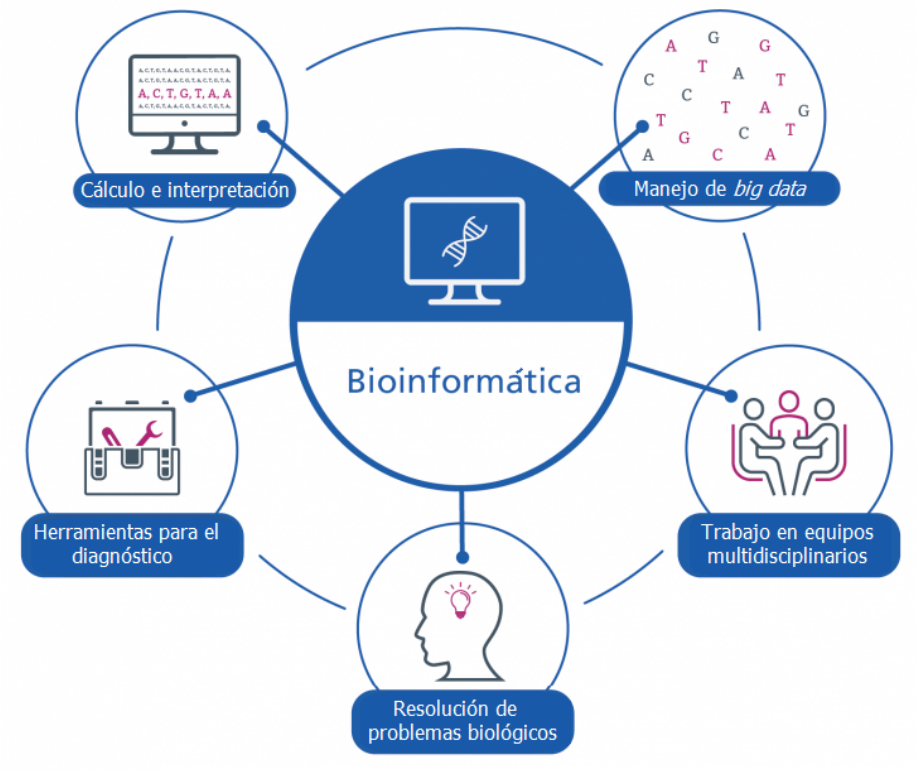

Figura 1.13: Características que forman parte de la Bioinformática.

En este sentido, se define a la Bioinformática como una subdisciplina de la biología y las ciencias de la computación que se encarga de adquirir, almacenar, analizar y diseminar la información biológica, en gran parte correspondiente a las secuencias de ADN y aminoácidos. La Bioinformática usa programas informáticos que tienen muchas aplicaciones, como por ejemplo: determinar las funciones de genes y proteínas, establecer relaciones evolutivas y predecir la conformación tridimensional de las proteínas. [National Human Genome Research Institute, 2020].

El gran salto de innovación tecnológica ha aumentado considerablemente la facilidad de producción y reducido significativamente el costo de recopilar a gran escala datos de miles de biomoléculas (genes, proteínas y metabolitos, principalmente). De esta manera, se permitió el crecimiento de la genómica, proteómica y metabolómica, todas áreas de estudio que se enfocan en el estudio de un sistema biológico de manera

holística [Horgan and Kenny, 2011] (Figura 1.14).

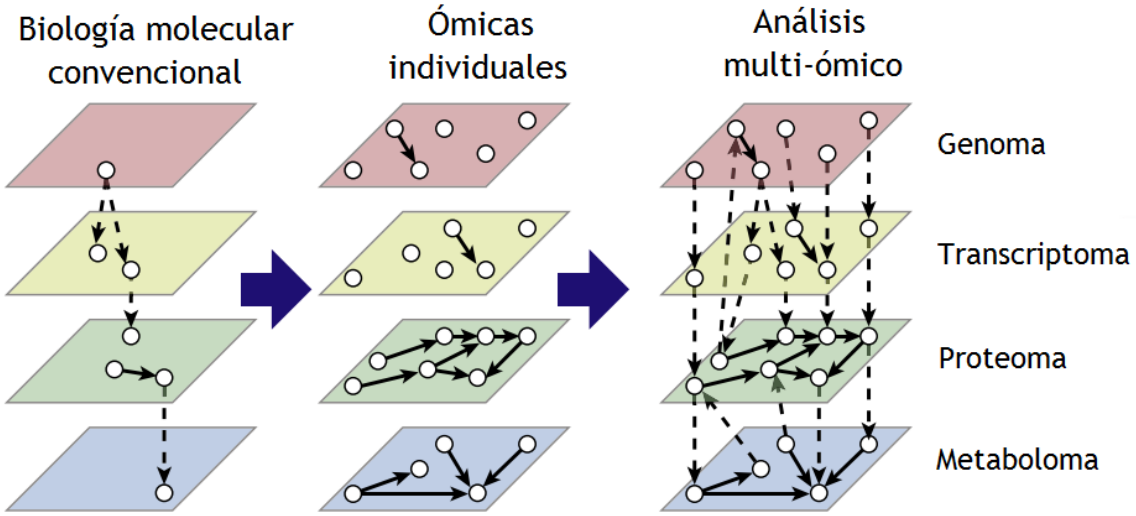

Figura 1.14: Un grupo de moléculas con propiedas químicas similares (genoma, proteoma, metaboloma) es una capa ómica en esta representación. La Bioinformática permite analizar y generar conocimiento biológico a partir de los distintos niveles de enfoque: mediante la acumulación de evidencia sobre moléculas específicas en la biología molecular convencional hasta la reconstrucción de redes de datos ómicos. Adaptado de [Yugi et al., 2016]

Estas tecnologías "ómicas"han tenido un gran impacto en el descubrimiento de métodos diagnósticos, biomarcadores y drogas, importantes en la era de la **Medicina de Precisión**, un concepto que amplía el conocimiento biológico y explora la gran diversidad de los individuos dentro de las características de una población, para brindar cuidados de salud personalizados.

El enfoque bioinformático mediante el uso de herramientas de aprendizaje automático permite una integración exhaustiva de datos clínicos multi-ómicos, interviniendo en etapas como el manejo y análisis de información genética, interpretación de la funcionalidad de genes, mecanismos de regulación celulares, selección y diseño de drogas *target*, identificación de enfermedades, entre otras.

1.3. Aprendizaje Automático

La generación de conocimiento biológico a partir del gran flujo de datos generados por las nuevas tecnologías en las ciencias biomédicas requiere el uso de herramientas y técnicas provenientes de las ciencias de la computación para complementar el análisis

bioinformático.

El Aprendizaje Automático o *Machine Learning* (ML) es una rama de la Inteligencia Artificial e incluye un conjunto de métodos que apuntan a que un sistema aprenda un modelo (hipótesis) a partir de una cantidad determinada de datos, tal que pueda identificar uno o más patrones contenidos en los mismos. Los patrones identificados (aprendidos) pueden ser usados para hacer estimaciones (predicciones) sobre datos nuevos similares a los datos usados para armar el modelo. En principio existen dos tipos principales de aprendizaje usado en ML: supervisado y no supervisado (Cuadro 1.1).

	Supervisado	**No Supervisado**
Datos	Pre-categorizados y cantidad conocida de clases	Sin categorizar y clases no conocidas
Objetivo	Predicción y Clasificación	Reconocimiento de patrones
Algoritmo	Clasificación y Regresión	*Clustering* y Asociación
Entrenamiento	Conjuntos de entrenamiento y evaluación	Solamente en base a propiedades intrínsecas de datos

Cuadro 1.1: Comparación de las principales características del aprendizaje supervisado y no supervisado.

Los métodos de aprendizaje **supervisado** generan modelos en base a datos de entrenamiento, los cuales representan características x_1, x_2,... ,x_p, evaluadas en n observaciones y una respuesta objetivo de salida Y, la cual es cuantificada en esas mismas observaciones. El aprendizaje supervisado formaliza una expresión que mapea los datos de entrenamiento y la respuesta *target*. Una vez que el sistema aprendió esta función, el modelo podrá predecir Y usando x_1, x_2, ... ,x_p ante la presentación de un nuevo dato.

Por otra parte, el aprendizaje **no supervisado** es una aproximación aplicable cuando la variable de respuesta resultante Y es desconocida. Se tiene solamente un conjunto de características x_1,x_2,...,x_p evaluadas en n observaciones, a partir de las

cuales el sistema aprende a hacer inferencias de acuerdo a similaridades y características intrínsecas. Dado que no se proporcionan etiquetas en los datos, no existe una forma específica de comparar el rendimiento del modelo en la mayoría de los métodos de aprendizaje no supervisados. Los principales algoritmos usados en cada tipo de aprendizaje están indicados en la Figura 1.15.

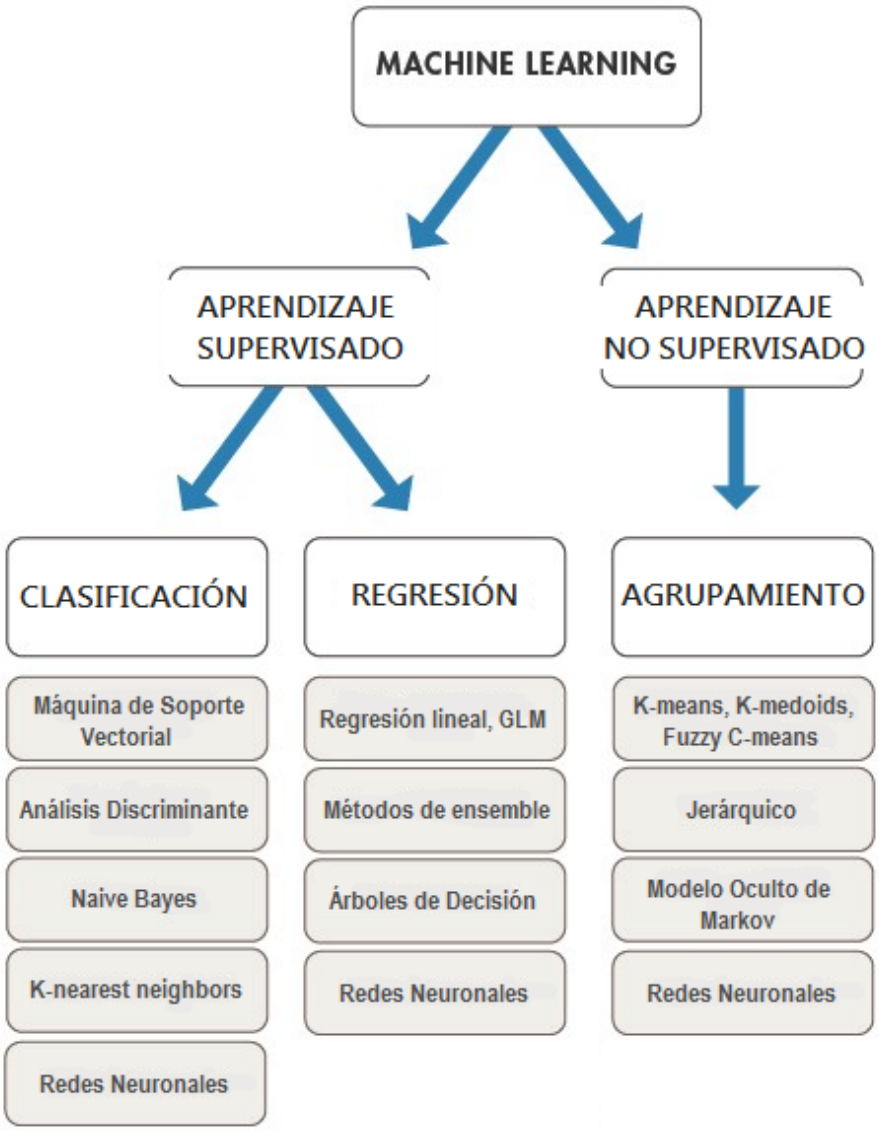

Figura 1.15: Principales algoritmos de ML.

Debido a su particular capacidad para manejar grandes y complejos conjuntos de datos y hacer predicciones sobre los mismos mediante la generación de modelos precisos, el uso de ML se expandió rápidamente en la comunidad de biociencias y en particular en el análisis bioinfomático [Lai et al., 2018, Shastry and Sanjay, 2020].

1.4. Objetivos y Sinopsis

El objetivo general de esta tesis es contribuir a la caracterización del microbioma intestinal humano en una subpoblación de individuos mediante el enfoque de un análisis de minería de datos y aprendizaje automático.

Por ello, los objetivos específicos son:

- Realizar el análisis bioinformático de los datos mediante el uso de algoritmos de aprendizaje automático, ampliando resultados previamente publicados mediante la búsqueda de potenciales asociaciones biológicas no descritas previamente.
- Determinar la importancia biomédica de las relaciones entre las variables demográficas de los pacientes y la información obtenida a partir del microbioma fecal.
- Seleccionar las herramientas de aprendizaje automático más útiles para analizar y visualizar los resultados de bioinformática.

La tesis está organizada de la siguiente manera:

- En el Capítulo 1 (página 25) se presentan los conceptos básicos de biología molecular y microbioma intestinal necesarios para establecer el marco teórico de esta tesis, además de aprendizaje automático.
- En el Capítulo 2 (página 47) se describen los materiales y métodos usados.
- En el Capítulo 3 (página 51) se realiza una descripción y análisis exploratorio del conjunto de datos.
- En el Capítulo 4 (página 61) se analiza la dimensionalidad mediante Análisis de Componentes Principales (PCA) y agrupamiento de los datos.
- En el Capítulo 5 (página 75) se implementa el método no supervisdo de Mapas Auto-Organizados (SOM).

- En los Capítulos 6 (página 101) y 7 (página 113) se implementan los algoritmos supervisados *Random Forest* y *Gradient Boosting Decision Tree*, respectivamente.

- En el Capítulo 8 (página 123) se discuten los resultados obtenidos y su significancia biomédica en el estudio del microbioma humano.

- Finalmente, en el Capítulo 9 (página 129) se resume brevemente los resultados y propone recomendaciones para análisis futuros.

Para llevar a cabo los objetivos planteados, se muestra a continuación un esquema de los pasos y métodos usados (Figura 1.16):

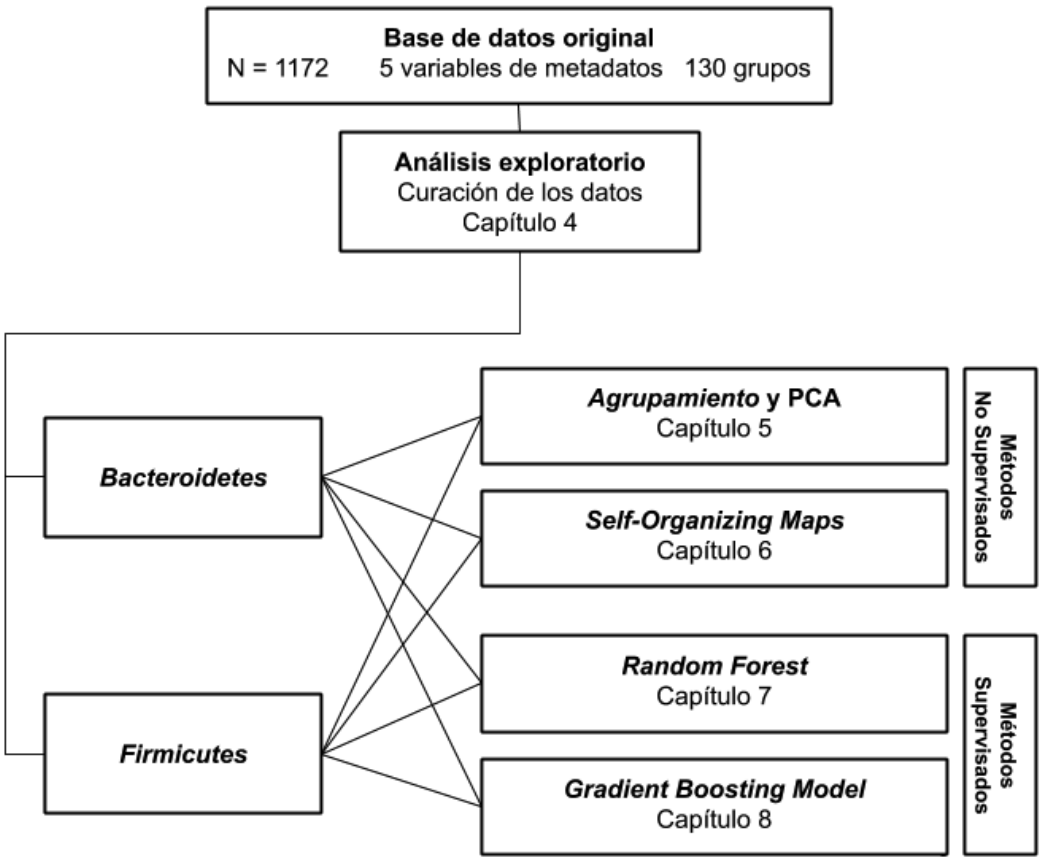

Figura 1.16: Diagrama de flujo que ilustra el flujo de trabajo general del enfoque metodológico en esta tesis

Capítulo 2

Materiales y Métodos

2.1. Conjunto de datos

Los datos utilizados en esta tesis fueron obtenidos del repositorio digital *DRYAD* [Dryad Digital Repository, 2021] que ofrece diversos tipos de datos curados para uso libre en investigación. El conjunto de datos usado incluye información de 1172 individuos adultos provenientes de 15 países occidentales agrupados en 6 regiones:

- Europa Central (Bélgica, Dinamarca, Alemania, Países Bajos).
- Europa del Este (Polonia).
- Países nórdicos (Finlandia, Noruega, Suecia).
- Europa del Sur (Francia, Italia, Serbia, España).
- Reino Unido/Irlanda.
- Estados Unidos.

.

A partir de muestras de materia fecal de los individuos, se determinó la abundancia relativa para 130 grupos de bacterias intestinales a través de las intensidades de señal del gen 16S rRNA, frecuentemente usado para la identificación de bacterias poco descriptas o no cultivables. Los valores de intensidad fueron generados mediante el

uso de *microarrays* filogenéticos *HITChip*. Estos tipos de *microarrays* ofrecen una plataforma de análisis sensible mediante la detección de las regiones hipervariables V1 y V6 del gen 16S rRNA, pudiendo detectar 1033 tipos de bacterias que representan la mayoría de la diversidad bacteriana en el intestino humano. Los 130 grupos de bacterias son referidas como especies y relacionados, apareciendo el término abreviado como .^{et} rel".

Los metadatos incluían el origen geográfico de los individuos (referido en el texto como "Nacionalidad"), categoría del índice de masa corporal (indicado como "BMI", por sus siglas en inglés), edad, sexo e índice de Shannon de diversidad bacteriana.

2.2. Procesamiento y limpieza

Los datos extraídos del repositorio público contenían algunos valores faltantes (NAs), lo cual podía representar un inconveniente al momento de aplicar algunos algoritmos. Se usaron las funciones de los paquetes *VIM* [Kowarik and Templ, 2016] y *tidyverse* [Wickham et al., 2019] en R para evidenciar los NAs. Las registros que contenían NAs fueron cuidadosamente removidos sin causar sesgo en el conjunto de datos. De esta manera, no se utilizó ningún método de imputación.

El conjunto de datos final estuvo representado por 1056 registros completos conteniendo datos de abundancia para 130 bacterias y metadatos de los individuos. Para los gráficos del análisis exploratorio se usaron las funciones del paquete *ggplot2* en R [Wickham, 2016].

2.3. Mapas auto-organizados (SOM)

La grilla en dos dimensiones del SOM puede ser implementada en diferentes topologías de n-filas x m-columnas.

Se seleccionó una topología hexagonal regular y se entrenó el SOM usando 750 neuronas, en grillas de 25 x 30.

Se usaron dos subconjuntos de datos para el entrenamiento, correspondientes a los

dos grupos mayoritarios de bacterias intestinales (*Bacteroidetes* y *Firmicutes*). Los datos fueron escalados usando escala logarítmica en base 10 antes de ser procesados en el SOM. El entrenamiento y la visualización del SOM fueron implementados mediante el paquete *kohonen* de R [Wehrens and Kruisselbrink, 2018].

Para visualizar los resultados del SOM, se usaron mapas de calor y de agrupamiento. Se utilizó agrupamiento jerárquico sobre el SOM entrenado.

Los mapas de calor fueron generados mediante el mapeo por separado de cada elemento en los pesos de un nodo sobre el mapa 2D. La escala de color fue configurada para que el color rojo represente el valor más alto en el mapa y el color azul represente el valor más bajo.

El número total de mapas de calor que se obtienen es igual a la cantidad de elementos que el peso de un nodo contiene (es decir, la cantidad de variables que tienen los datos de entrenamiento).

2.4. Algoritmos basados en árboles de decisión

Los datos de abundancia de bacterias y los metadatos fueron usados como regresores para generar un modelo de predicción de la diversidad usando *Random Forest* y *Gradient Boosting Decision Trees*, implementados mediante las herramientas DRF (*Distributed Random Forest*) y GBM (*Gradient Boosting Machine*) provistas por la interfase R de *h2o.ai* [h2o, 2020].

Se usaron dos subconjuntos de datos para el entrenamiento, correspondientes a los dos grupos mayoritarios de bacterias intestinales (*Bacteroidetes* y *Firmicutes*).

En el entrenamiento de los modelos de regresión se consideró *cross-validation*=10 y el desempeño fue evaluado considerando métricas de error como el error absoluto medio (MAE). Se calibró un número variable de árboles, luego del cual se seleccionó el mejor modelo de regresión en ambos métodos.

Capítulo 3

Análisis exploratorio de la población en estudio

El trabajo original que inspiró el uso de los datos para la realización de la presente tesis estudió la distribución de bacterias a lo largo de gradientes fluídos o mediante cambios abruptos entre estados estables alternativos, una propiedad característica de un sistema ecológico [Lahti et al., 2014]. En base a este concepto, los autores observaron la existencia de dos estadios contrastantes de abundancia para ciertas poblaciones de bacterias: alta o baja abundancia, mostrando una estabilidad reducida entre estas dos fases. Para tener otro enfoque de análisis sobre el conjunto de datos, a continuación se describirán las características de los datos y un análisis exploratorio de los mismos.

3.1. Identificación de valores faltantes

El problema de la existencia de datos faltantes en conjuntos de datos reales es relativamente común y puede tener un efecto significativo en las conclusiones a extraer. El análisis inicial del conjunto de datos total mostró que las variables correspondientes a metadatos fueron las únicas que presentaron valores faltantes (NAs) y no así las variables correspondientes a los niveles de abundancia para cada bacteria (Figura 3.1).

La variable BMI acumuló un 9 % de NAs, seguido por Edad, Sexo y Nacionalidad (Cuadro 3.1).

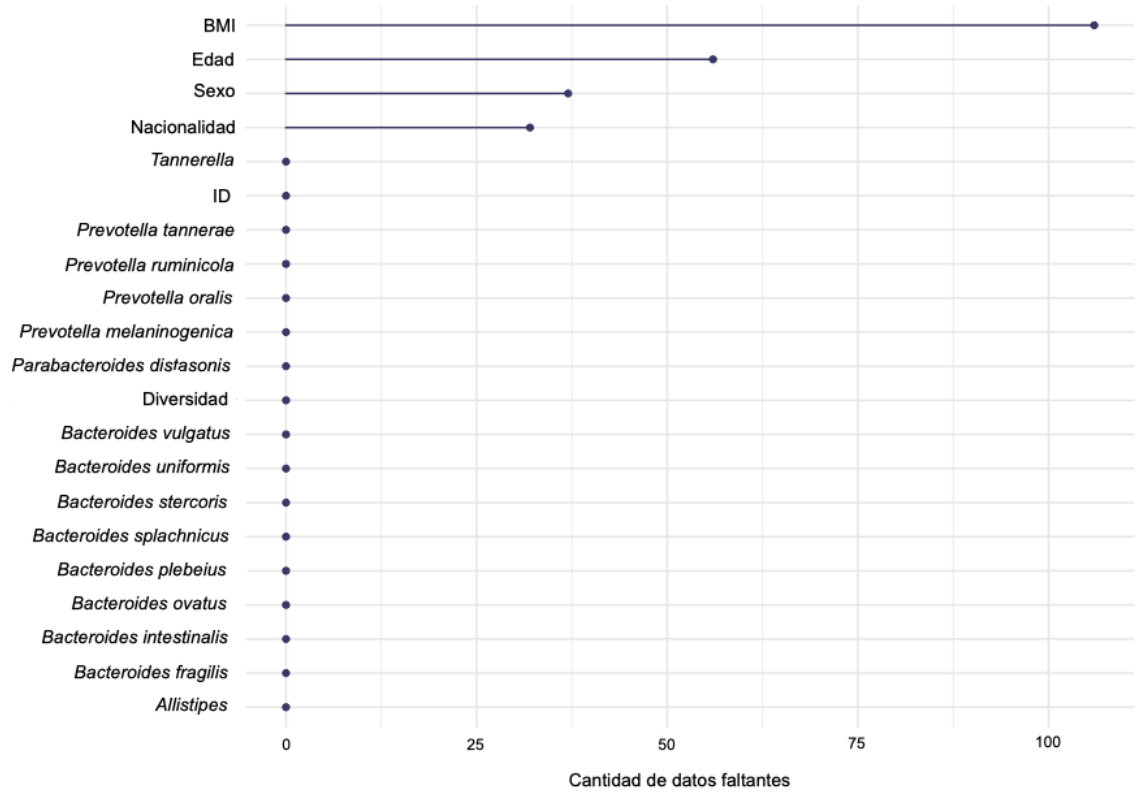

Figura 3.1: Cantidad de valores faltantes en el conjunto de datos total. De las 130 poblaciones de bacterias, ninguna presentó valores faltantes. Se muestran solamente una cantidad pequeña de bacterias.

Variable	Cantidad	%
BMI	106	9
Edad	56	4.7
Sexo	37	3.1
Nacionalidad	32	2.7

Cuadro 3.1: Variables del conjunto de datos original que presentan valores faltantes.

El patrón de distribución de los valores faltantes en las variables de metadatos se combinó de distintas maneras. De un total de 1172 registros, 49 de ellos tenían valores vacíos solamente en la variable BMI; 22 casos presentaron valores faltantes en combinación con BMI, Edad, Sexo y Nacionalidad; 19 casos presentaron valores

faltantes combinados en BMI y Edad; 15 casos faltantes de manera conjunta entre BMI, Edad y Sexo; 9 casos presentaron valores faltantes en la variable Nacionalidad solamente y 1 caso presentó valor faltante de manera conjunta en las variables BMI y Nacionalidad. En total, la cantidad de registros que presentaron valores faltantes en el conjunto de datos fue 115 (Figura 3.2).

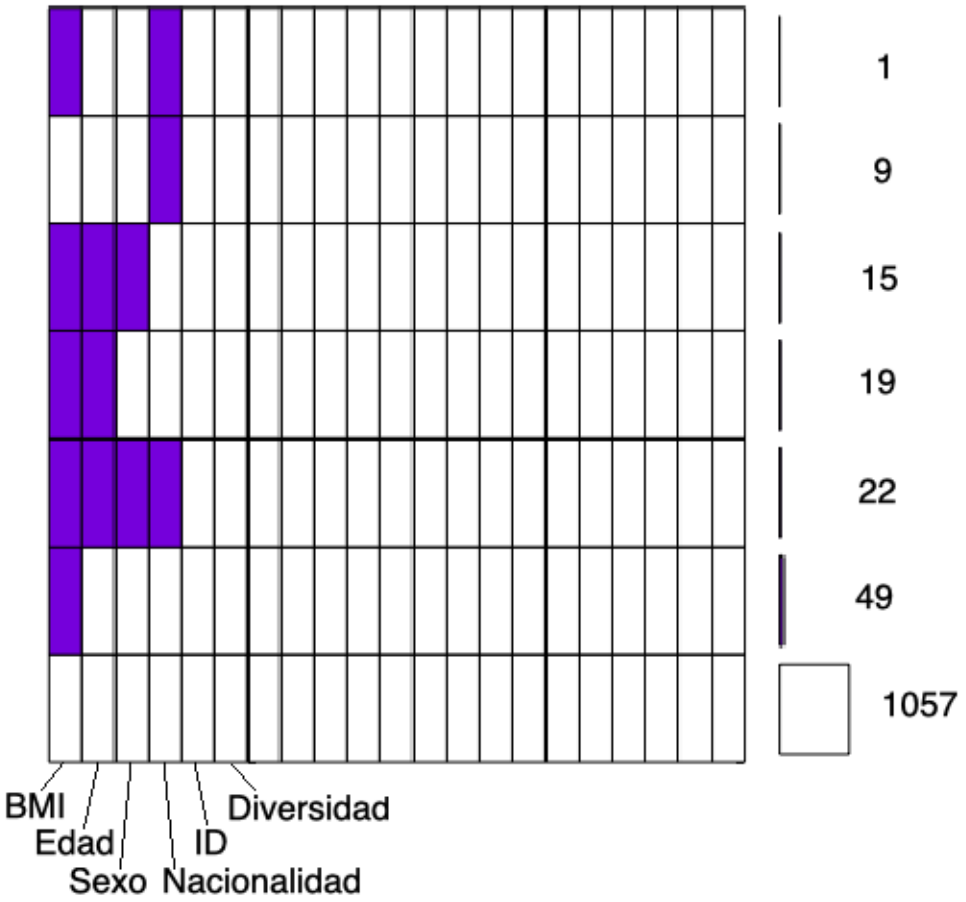

Figura 3.2: Combinaciones de las cuatro variables de metadatos con valores faltantes (indicados en violeta). Las variables completas están indicadas sin color. Las celdas sin nombre corresponden a un subconjunto de variables de bacterias.

Teniendo en cuenta que el conjunto de datos fue tomado de un repositorio público y que no existe información adicional sobre el modo en el cual fue recopilada la información o indicaciones que justifiquen la presencia de algunos campos vacíos, no es posible determinar la naturaleza del tipo de dato faltante en los datos para así usar algún método de imputación apropiado. De esta manera, se decidió no tener en cuenta a esta cantidad de valores NA en los posteriores análisis y trabajar con un conjunto

de datos completo.

Una manera de controlar que el impacto de la eliminación de N=115 registros faltantes no afecte al conjunto total, es decir que no produzca ningún sesgo, es mediante la observación del efecto de eliminación de los NAs en las variables. La eliminación de los 115 registros vacíos no alteró la distribución de las variables BMI, Sexo y Nacionalidad, a excepción de la categoría Europa del Este que quedó representado solamente por un único registro completo (Cuadro 3.2).

Categoría	N original	N sin NAs
Nacionalidad		
Europa Central	650	617
Europa del Este	15	1
Escandinavia	291	275
Europa del Sur	90	90
Reino Unido/Irlanda	50	40
Estados Unidos	44	34
NAs	32	0
BMI		
Delgado	493	487
Mórbido obeso	22	22
Obeso	224	224
Sobrepeso	204	201
Obeso severo	100	100
Bajo peso	23	23
Sexo		
Femenino	680	646
Masculino	455	411

Cuadro 3.2: Distribución de individuos antes (N original) y después (N sin NAs) de la eliminación de los registros que presentaban campos vacíos.

De esta manera, en los análisis posteriores no se incluirá al único registro remanente de Europa del Este dado que no es representativo. Entonces, en los siguientes capítulos se considerará un conjunto de datos de N=1056 registros completos.

3.2. Metadatos

Cuatro regiones europeas están representadas en el conjunto de datos: Europa del Sur y Central abarcan de manera extensa todo el espectro de edades observado en el conjunto de datos, Escandinavia tiende a concentrar individuos de edad media (alrededor de 50 años), mientras que UK/Irlanda concentra individuos jóvenes en su mayoría (entre 20 y 30 años). Por otra parte, en este conjunto de datos existe representación de una población fuera de Europa: Estados Unidos (US) abarca también edades jóvenes y adultas (Figura 3.3).

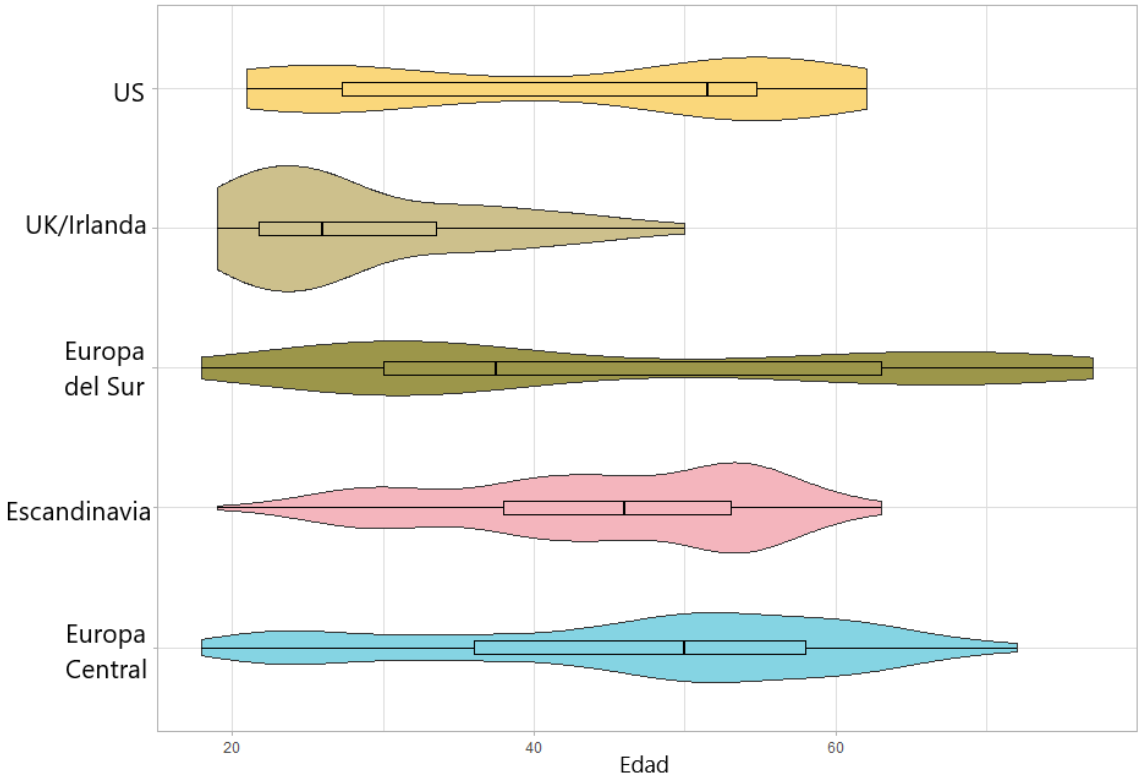

Figura 3.3: Distribución de las franjas etarias en las distintas regiones geográficas del conjunto de datos.

Es interesante destacar que el conjunto de datos incluye además información sobre una variable fisiológica básica como es el índice de masa corporal (BMI, por sus siglas en inglés). El BMI es una medida que indica el estado nutricional de un individuo y se define por el peso (en kg) dividido por el cuadrado de la altura (en metros). Los rangos de valores de BMI están basados en el efecto que la grasa corporal en exceso tiene sobre el riesgo de enfermedad: a medida que aumenta el BMI, también aumenta

el riesgo de algunas enfermedades [World Health Organization, 2021].

En este conjunto de datos los individuos delgados (BMI = 18.5-25) se distribuyeron de manera homogénea en todas las franjas etarias, mientras que los individuos con sobrepeso (BMI = 25-30), obesos (BMI = 30-35) y obeso severos (BMI = 35-40) se concentraron entre los 45-60 años. Una gran proporción de población de bajo peso estuvo representada entre los 20-30 años de edad (Figura 3.4).

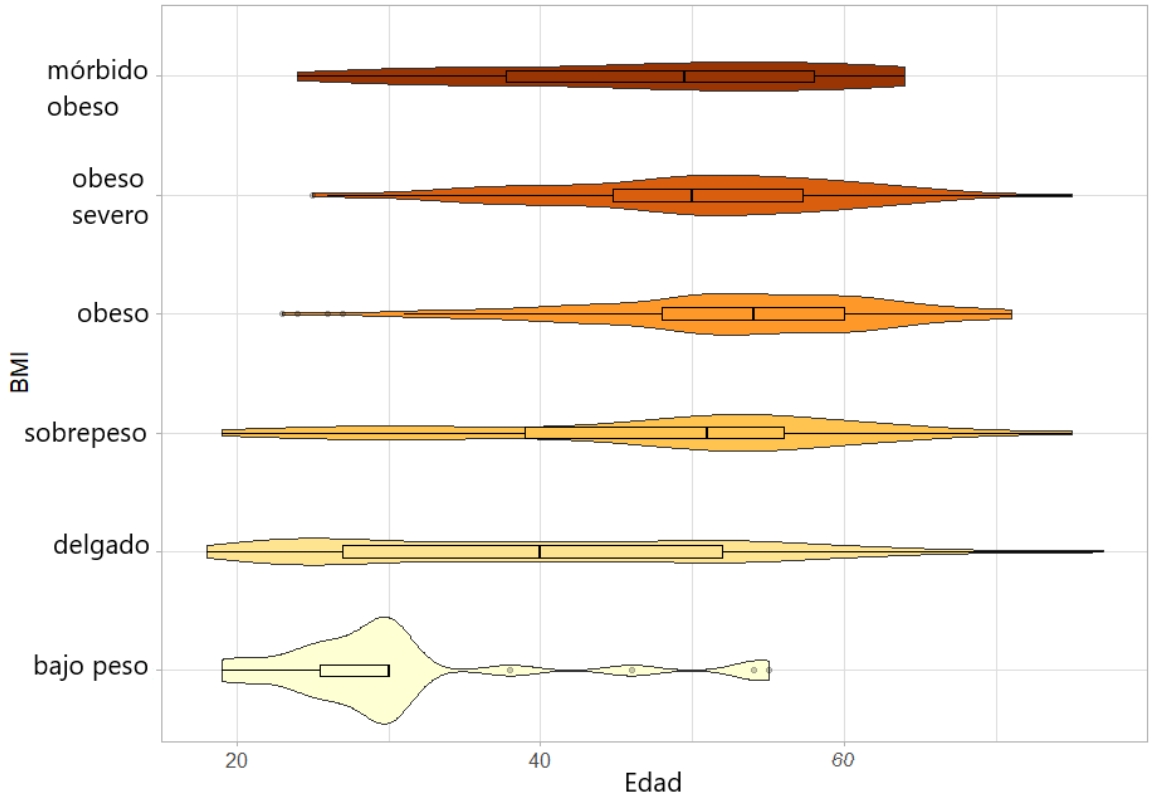

Figura 3.4: Distribución de las categorías de BMI en función de la edad.

La distribución de las diferentes categorías de BMI en cada región geográfica mostró una representación predominante de individuos delgados y proporciones variables de los restantes niveles de BMI. Tanto en Europa del Sur como en UK/Irlanda, los individuos delgados representaron un poco más de la mitad de esas subpoblaciones, seguidos en cantidad por individuos con sobrepeso. En cambio, en Europa Central, Escandinavia y US, la proporción de individuos delgados representó un poco menos de la mitad de esas subpoblaciones, seguidas por individuos obesos. En Europa Central se observó representación variable de todas las categorías de BMI (Figura 3.5).

El conjunto de datos incluye además una variable que representa la diversidad en

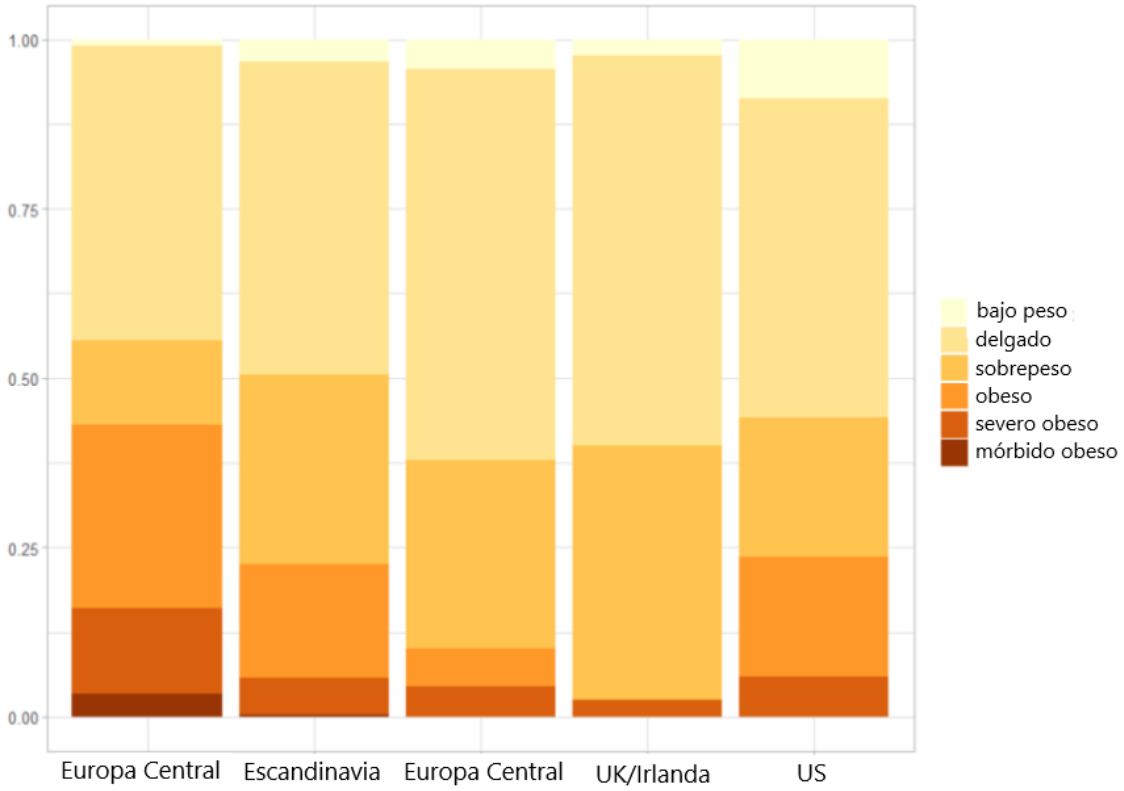

Figura 3.5: Proporción de índices de masa corporal (BMI) de acuerdo a la región geográfica.

términos del índice de Shannon. En el análisis de comunidades ecológicas, la diversidad es una propiedad ampliamente estudiada porque tiene en cuenta la riqueza (la cantidad de clases) y uniformidad (la distribución de individuos entre las clases) de los miembros. El entendimiento de la diversidad en la comunidad microbiana intestinal permite además comprender el impacto de diversos factores tales como el uso de antibióticos, tipo de dieta, grado de obesidad, intervenciones médicas y factores ambientales, entre otros (revisado en [Reese and Dunn, 2018]). De las seis categorías de BMI presentes en el conjunto de datos, los individuos delgados mostraron el mayor valor de diversidad mientras que la población mórdibo-obesa tuvo el valor más bajo de diversidad (Figura 3.6).

3.3. Selección de grupos de bacterias

La clasificación taxonómica de bacterias está representada por varios grupos principales (denominados *phyla*) que engloban una o más clases, los que a su vez contienen

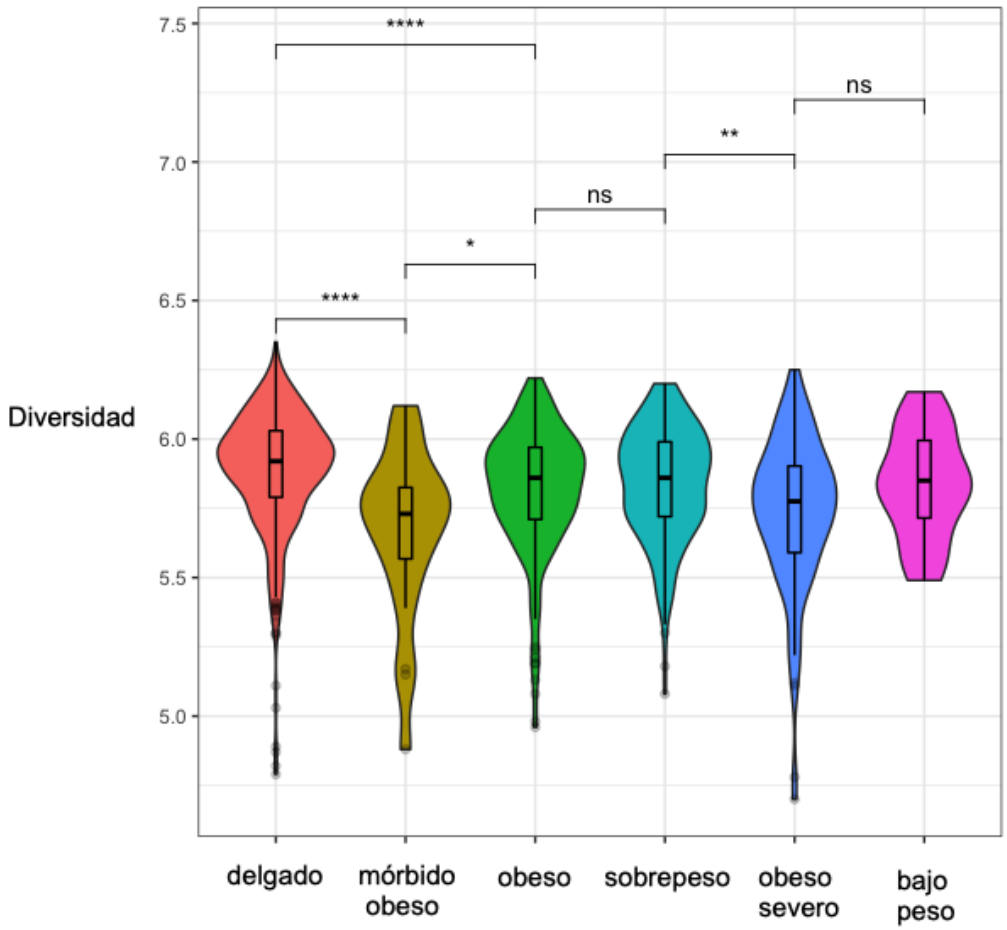

Figura 3.6: Distribución de la Diversidad total del conjunto de datos en relación a los grupos de BMI.

órdenes, familias, géneros y especies. Los *phyla* predominantes en el microbioma intestinal humano son: *Bacteroidetes*, *Firmicutes*, *Actinobacteria*, *Proteobacteria*, *Fusobacteria* y *Verrucomicrobia*. Los dos grupos que representan el 90 % de la microbiota intestinal son *Bacteroidetes* y *Firmicutes* [Arumugam et al., 2011].

La inspección inicial del conjunto de datos indicó que de los 130 tipos de bacterias caracterizadas por el *microarray* filogenético, 8 miembros pertenecen al *phylum Actinobacteria*, mientras que los *phyla Fusobacteria* y *Verrucomicrobia* están representados por un único miembro cada uno. Se observó además que 15 tipos de bacterias pertenecen al *Bacteroidetes*, con predominancia de los géneros *Bacteroides* y *Prevotella* (Cuadro 3.3). En cuanto a *Firmicutes*, se identificaron 73 miembros (Cuadro 3.4). Los restantes 32 tipos pertenecen al *phylum Proteobacteria*.

Teniendo en cuenta que los miembros de *Proteobacteria* tienen baja abundancia en

el instestino humano saludable [Shin et al., 2015] y que, en cambio, predominan las bacterias pertenecientes a *Firmicutes* y *Bacteroidetes*, se decidió enfocar los siguientes análisis usando los datos de estos dos grandes grupos.

En resumen, los capítulos siguientes usarán las variables de metadatos y la información de 15 tipos de bacterias de *Bacteroidetes* y 74 tipos de bacterias de *Firmicutes*, ambos para una población de individuos sanos de N=1056.

Especie / Género		
Bacteroides ovatus	*Bacteroides plebeius*	*Bacteroides stercoris*
Bacteroides splachnicus	*Bacteroides vulgatus*	*Bacteroides uniformis*
Bacteroides intestinalis	*Bacteroides fragilis*	*Allistipes sp.*
Parabacteroides distasonis	*Tannerella sp.*	*Prevotella tannerae*
Prevotella oralis	*Prevotella melaninogenica*	*Prevotella ruminicola*

Cuadro 3.3: Miembros del subconjunto *Bacteroidetes* presentes en los datos.

Especie / Género		
Aerococcus sp.	*Anaerofustis*	*Anaerostipes caccae*
Anaerotruncus colihominis	*Anaerovorax odorimutans*	*Aneurinibacillus*
Asteroleplasma	*Bacillus sp.*	*Bryantella formatexigens*
Bulleidia moorei	*Butyrivibrio crossotus*	*Catenibacterium mitsuokai*
Clostridium leptum	*Clostridium sensu stricto*	*Clostridium ramosum*
Clostridium stercorarium	*Clostridium cellulosi*	*Clostridium nexile*
Clostridium orbiscindens	*Clostridium symbiosum*	*Clostridium difficile*
Clostridium thermocellum	*Clostridium felsineum*	*Clostridium colinum*
Clostridium sphenoides	*Coprobacillus catenaformis*	*Coprococcus eutactus*
Dialister	*Dorea formicigenerans*	*Enterococcus*
Eubacterium biforme	*Eubacterium rectale*	*Eubacterium limosum*
Eubacterium siraeum	*Eubacterium hallii*	*E. cylindroides*
Eubacterium ventriosum	*Faecalibacterium prausnitzii*	*Gemella*
Granulicatella	*Lachnobacillus bovis*	*Lachnospira pectinoschiza*
Lactobacillus gasseri	*Lactobacillus catenaformis*	*Lactobacillus salivarius*
Lactobacillus plantarum	*Lactococcus*	*Megamonas hypermegale*
Megasphaera elsdenii	*Mitsuokella multiacida*	*Oscillospira guillermondii*
O. Clostridium XIVa	*Papillibacter cinnamivorans*	*Peptococcus niger*
Peptostreptococcus anaerobius	*Peptostreptococcus micros*	*P. faecium*
Roseburia intestinalis	*Ruminococcus lactaris*	*Ruminococcus gnavus*
Ruminococcus bromii	*Ruminococcus obeum*	*Ruminococcus callidus*
Sporobacter termitidis	*Staphylococcus*	*Streptococcus intermedius*
Streptococcus mitis	*Streptococcus bovis*	*Subdoligranulum variable*
UCI II	*Uncultured Selenomonadaceae*	*Veillonella*
Weissella		

Cuadro 3.4: Miembros del subconjunto *Firmicutes* presentes en los datos.

Capítulo 4

Reducción de la dimensionalidad y agrupamiento

Uno de los primeros pasos en el análisis de un conjunto de datos de dimensiones considerables es caracterizar al mismo en cuanto a la composición y estructura de sus variables. En principio, un conjunto de datos multivariado puede incluir la posiblidad de que la mayoría de las variables estén correlacionadas, lo que implica redundancia de información. Además, trabajar con todas las variables podría devolver un modelo no del todo preciso.

Muchos conjuntos de datos pueden ser vistos como una gran matriz, la cual puede ser resumida mediante la búsqueda de matrices más pequeñas (con menor cantidad de filas o columnas) que están relacionadas con la original. El manejo de matrices más pequeñas puede resultar en un análisis más eficiente en comparación con la gran matriz original. El proceso de calcular estas matrices más pequeñas es lo que se conoce como reducción de la dimensionalidad ([Leskovec et al., 2014]).

A continuación se presentarán los resultados de Análisis de Componentes Principales (PCA) y Agrupamiento (*clustering*), dos enfoques de análisis no supervisados sobre los subconjuntos *Bacteroidetes* y *Firmicutes*.

4.1. Análisis de Componentes Principales

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) tiene como objetivo principal reducir la dimensionalidad de variables de un conjunto de datos, tratando a su vez de conservar la información original.

La generación de un sistema de componentes principales hace que las muestras se distribuyan en un nuevo espacio de componentes, el cual es simplemente una rotación del espacio de las variables originales, manteniendo cualquier relación presente. Así, es posible visualizar las relaciones entre las muestras en base a todas las características además de observar la dirección de esas relaciones.

Existen varios criterios para determinar el número de componentes principales (PCs) a considerar en el análisis de PCA:

- Un criterio ampliamente usado es el que tiene en cuenta un número determinado de componentes principales que explican un porcentaje de interés de la variabilidad original. La elección del porcentaje es arbitraria, dependiendo de si se apunta a la reducción de la dimensionalidad o un análisis exploratorio con PCA.

- Otro criterio a considerar es el gráfico de los autovalores de cada componente (conocido como "gráfico de sedimentación"), el cual muestra cuánta variación es capturada por los mismos. Idealmente, es posible determinar un punto de corte para la cantidad de componentes a elegir si la curva muestra un descenso empinado y luego dobla en un “codo” a partir del cual se aplana.

4.1.1. *Bacteroidetes*

Los resultados indicaron que para explicar al menos un porcentaje significativo de variabilidad total en el subconjunto de *Bacteroidetes* es necesario considerar al menos 9 PCs, los cuales explican alrededor del 72.5 % de la varianza acumulada del subconjunto de datos (Cuadro 4.1).

PC	Proporción de varianza	Varianza acumulada
1	19,00	19,00
2	10,44	29,44
3	7,04	36,48
4	6,52	43,00
5	6,28	49,29
6	6,11	55,41
7	5,97	61,39
8	5,62	67,01
9	5,49	72,51
10	5,32	77,83
11	5,17	83,01
12	4,88	87,89
13	4,69	92,59
14	4,37	96,96
15	3,03	100

Cuadro 4.1: Aporte de cada componente principal a la varianza total en el subconjunto *Bacteroidetes*.

En cuanto al gráfico de sedimentación, se observó un punto de corte en el tercer componente como determinante (Figura 4.1).

El análisis exploratorio de PCA puede ser enfocado desde el punto de vista de la contribución de cada variable sobre los componentes principales examinando la magnitud y la dirección de los coeficientes de las variables originales o *loadings*. Cuanto mayor sea el valor absoluto del coeficiente, más importante será la variable correspondiente en el cálculo del componente.

A modo de un primer enfoque exploratorio del subconjunto de *Bacteroidetes*, se describirá la influencia de las variables sobre los cuatro primeros componentes solamente (Cuadro 4.2).

Todas las bacterias de este subconjunto de datos correlacionan de manera positiva con la varianza explicada por el PC1. Específicamente, se observó que todas las bacterias del género *Bacteroides* (además del género *Allistipes* y *Tannerella*) mostraron un alto grado de correlación, siendo *Bacteroides vulgatus* la de mayor contribución a la varianza del PC1. Por otra parte, 3 de las 4 especies de *Prevotella*, (a excepción de *P. tannerae*), mostraron una contribución casi nula a la varianza de este componente, siendo *P. ruminicola* la de menor contribución, seguida por *Prevotella melaninogenica* y *Prevotella oralis*. En base a lo mencionado, se podría interpretar entonces que **PC1**

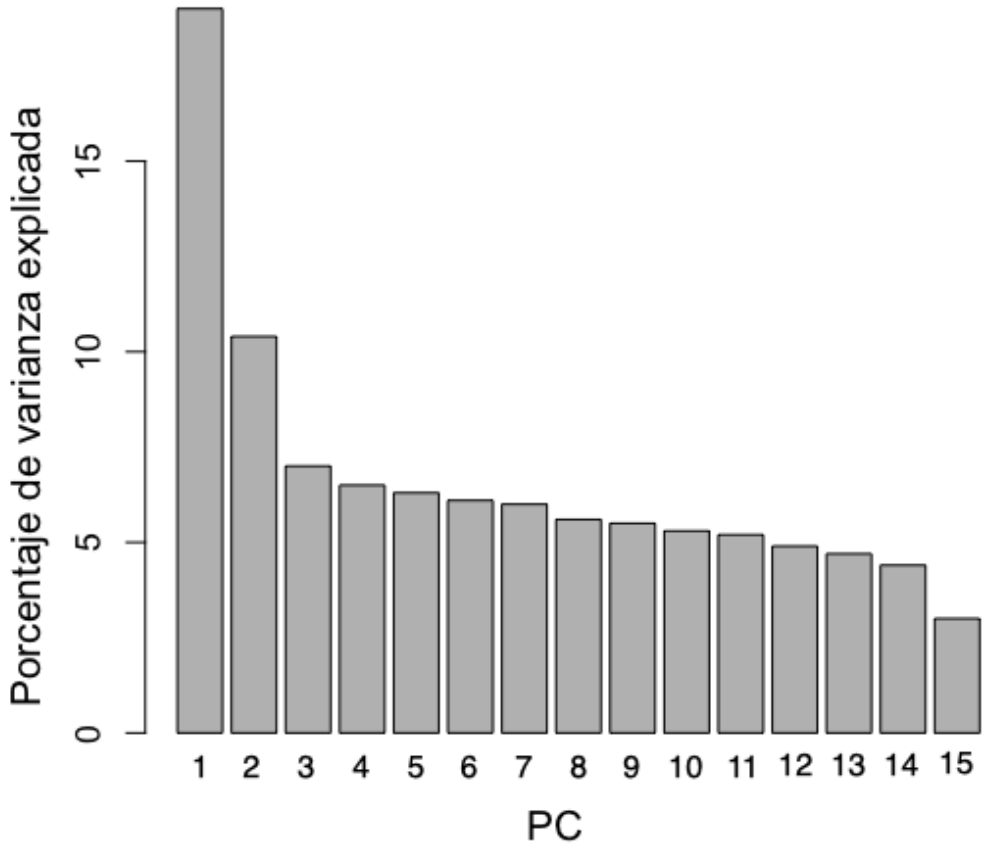

Figura 4.1: Gráfico de sedimentación de los autovalores de cada componente principal (PC) para el subconjunto de *Bacteroidetes*.

Variable	PC1	PC2	PC3	PC4
Allistipes	0,311	-0,041	0,007	-0,314
Bacteroides fragilis	0,328	-0,059	0,071	0,036
Bacteroides intestinalis	0,235	-0,133	0,261	-0,016
Bacteroides ovatus	0,255	-0,049	0,074	0,109
Bacteroides plebeius	0,247	0,082	-0,296	**-0,474**
Bacteroides splachnicus	0,238	0,021	**0,353**	0,369
Bacteroides stercoris	0,198	-0,010	0,004	**0,415**
Bacteroides uniformis	0,328	**-0,148**	-0,122	-0,310
Bacteroides vulgatus	**0,374**	0,001	-0,168	-0,078
Parabacteroides distasonis	0,275	0,043	0,204	0,084
Prevotella melaninogenica	0,049	**0,684**	0,064	-0,037
Prevotella oralis	0,086	0,678	0,112	-0,054
Prevotella ruminicola	**0,034**	0,115	**-0,762**	0,379
Prevotella tannerae	0,316	-0,005	-0,155	0,303
Tannerella	0,285	-0,027	-0,027	0,050

Cuadro 4.2: Importancias relativas de las variables en los cuatro primeros componentes principales de *Bacteroidetes*. En negrita se indican los valores máximos y mínimos obtenidos para cada PC.

discrimina entre los dos géneros más relevantes, *Bacteroides* y *Prevotella*, dentro del subconjunto *Bacteroidetes*.

Para PC2, se observó que un gran grupo de bacterias del género *Bacteroides* tuvieron una contribución casi nula y negativa, siendo *B. uniformis* la de mayor contribución negativa sobre la varianza de esta componente. En cambio, un grupo de bacterias del género *Prevotella* tuvieron una importante correlación positiva a la varianza de PC2 (*P. melaninogenica*, *P. oralis* y *P. ruminicola*). En base a esto, se podría interpretar a **PC2 como una dimensión del género *Prevotella***.

Por otra parte, para PC3 se observó una máxima correlación positiva para *Bacteroides splachnicus* y una correlación máxima negativa para *Prevotella ruminicola*. Estos resultados son relevantes dado que está descripto que ambas especies de bacterias son abundantes en una muestra de individuos no-obesos en comparación con obesos [Verdam et al., 2013]. Teniendo en cuenta esto, podría interpretarse al **PC3 como discriminante de dos perfiles de bacterias relevantes en obesidad**.

Toda esta información puede ser visualizada en un *biplot*, el cual representa las muestras individuales junto a vectores que corresponden a los *loadings*, cuya longitud y dirección indica la influencia de cada variable en el espacio del componente, PC1-PC2 (Figura 4.2) y PC3-PC4 (Figura 4.3).

4.1.2. *Firmicutes*

Los resultados indicaron que para explicar al menos un porcentaje significativo de varianza total en el subconjunto de *Firmicutes* es necesario considerar 42 PCs que representan alrededor del 70 % de la varianza acumulada del subconjunto de datos (Cuadro 4.3).

Así mismo, el gráfico de sedimentación mostró una curva descendiente gradual que no permite considerar ningún componente en particular como punto de corte (Figura 4.4).

Teniendo en cuenta los resultados mencionados, es posible reducir la dimensionalidad del subconjunto *Firmicutes* a 42-43 componentes. Pero dado que el porcentaje

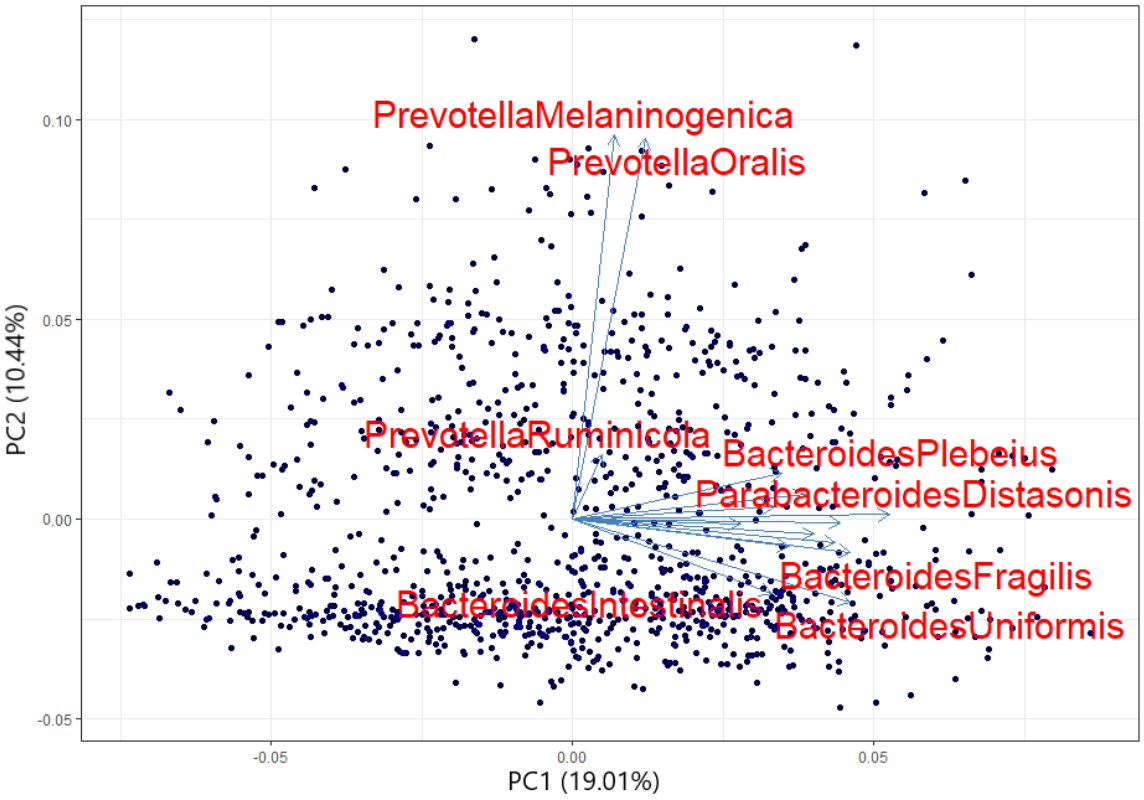

Figura 4.2: *Biplot* representando la posición de cada muestra en términos de PC1 y PC2 para *Bacteroidetes*.

de varianza explicado por los primeros componentes es muy bajo, el enfoque exploratorio de PCA sobre este subconjunto se torna forzado, dada la baja contribución a la varianza. En la visualización del *biplot* para PC1 y PC2 se observó una distribución de los datos concentrada y sin *clusters* aparentes. Debido a esto y para obtener una mejor visualización, los vectores de dirección de los *loadings* fueron omitidos a propósito (Figura 4.5).

4.2. Agrupamiento - *K-means*

El agrupamiento, o *clustering*, se define como una clasificación no supervisada de un conjunto de datos, apuntando a que los objetos dentro de un subgrupo sean lo más similar posible (es decir, una alta similitud intra-clase) mientras que los objetos de diferentes subgrupos sean lo más diferente entre sí (baja similitud entre-clase).

El componente de aprendizaje no supervisado implica que no hay una variable de respuesta para entrenar y encontrar relaciones entre las observaciones. En el flujo de

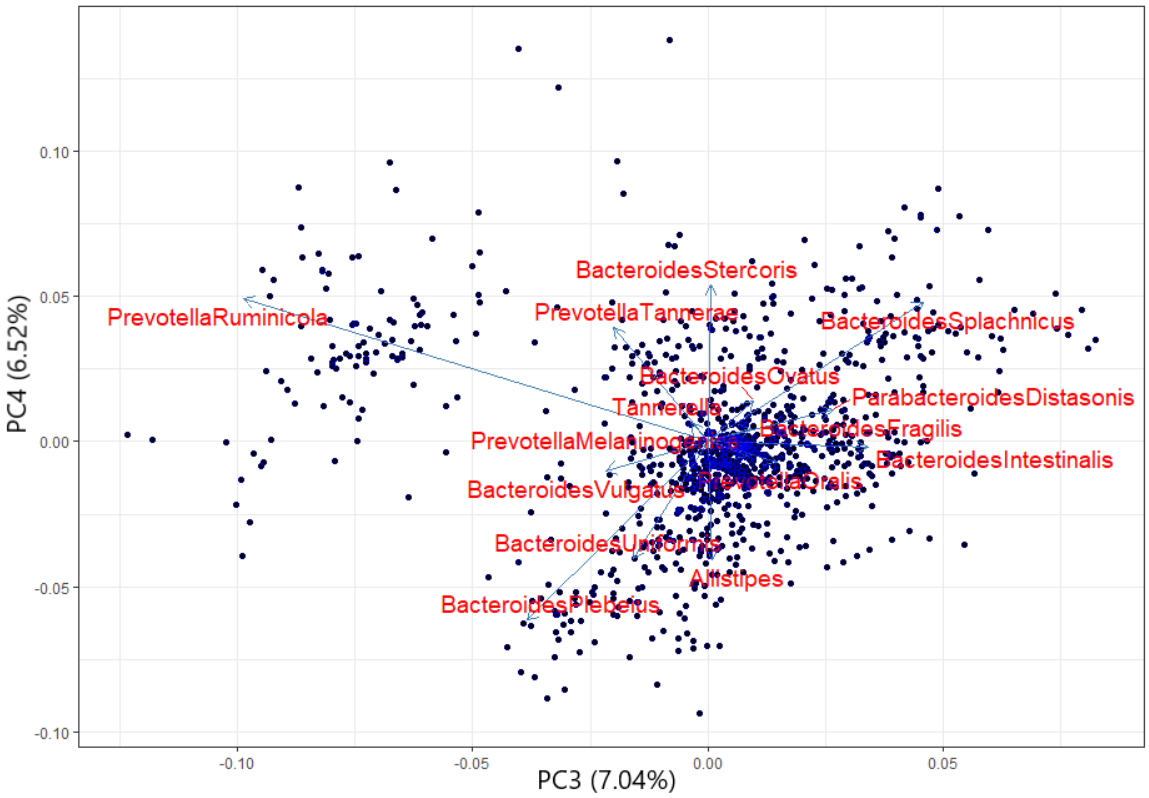

Figura 4.3: *Biplot* representando la posición de cada muestra en términos de PC3 y PC4 para *Bacteroidetes*.

análisis de un conjunto de datos, el uso de un determinado algoritmo de *clustering* complementa análisis posteriores para un mejor entendimiento de la naturaleza de los datos.

K-means es un método de partición de los datos en un conjunto de k grupos o *clusters*. Cada *cluster* está representado por sus centroide (el valor medio de las observaciones). *K-means* usa una mejora iterativa para producir un resultado final comenzando con estimaciones iniciales para los k centroides seleccionados al azar y así la iteración ocurre entre dos pasos:

- Asignación de los datos: Cada dato es asignado a su centroide más cercano, en base a una medida de distancia o similaridad.
- Actualización del centroide: En este paso, se recalculan los centroides mediante el promedio de todos los datos asignados a cada *cluster*.

El algoritmo se detiene cuando los centroides se han estabilizado (es decir, no hay más cambio que se obtenga de agrupamientos previos) y se ha alcanzado el número

PC	Proporción de varianza	Varianza acumulada
1	0,040	0,040
2	0,026	0,067
3	0,021	0,089
4	0,020	0,109
5	0,020	0,129
6	0,019	0,149
7	0,019	0,168
8	0,018	0,187
9	0,018	0,205
10	0,017	0,223
...	...	...
40	0,012	0,666
41	0,012	0,679
42	0,012	0,691
43	0,011	0,702
44	0,011	0,714
45	0,011	0,726
...	...	...
70	0,007	0,970
71	0,007	0,977
72	0,007	0,985
73	0,007	0,992
74	0,007	1,000

Cuadro 4.3: Varianza total explicada por cada componente principal del subconjunto *Firmicutes*.

definido de iteraciones.

El valor de k debe ser especificado con anticipación y su determinación depende en general del conocimiento del dominio y de los datos, entre otros factores. Existen métodos directos y de evaluación estadística que ayudan a determinar el valor de k y se enfocan en la optimización de un criterio. Entre los métodos directos se pueden mencionar a:

- Método del codo o *elbow method*, mediante la minimización del valor de *WSS* (*Within cluster Sums of Squares*), es decir, la suma de cuadrados dentro del *cluster*.

- Coeficiente de *Silhouette*, el cual evalúa la calidad del *cluster* y estima las distancias de cada punto intra e inter *cluster*, mediante la similitud y disimilitud.

La idea principal en estos métodos es lograr una máxima cohesión de los *clusters*,

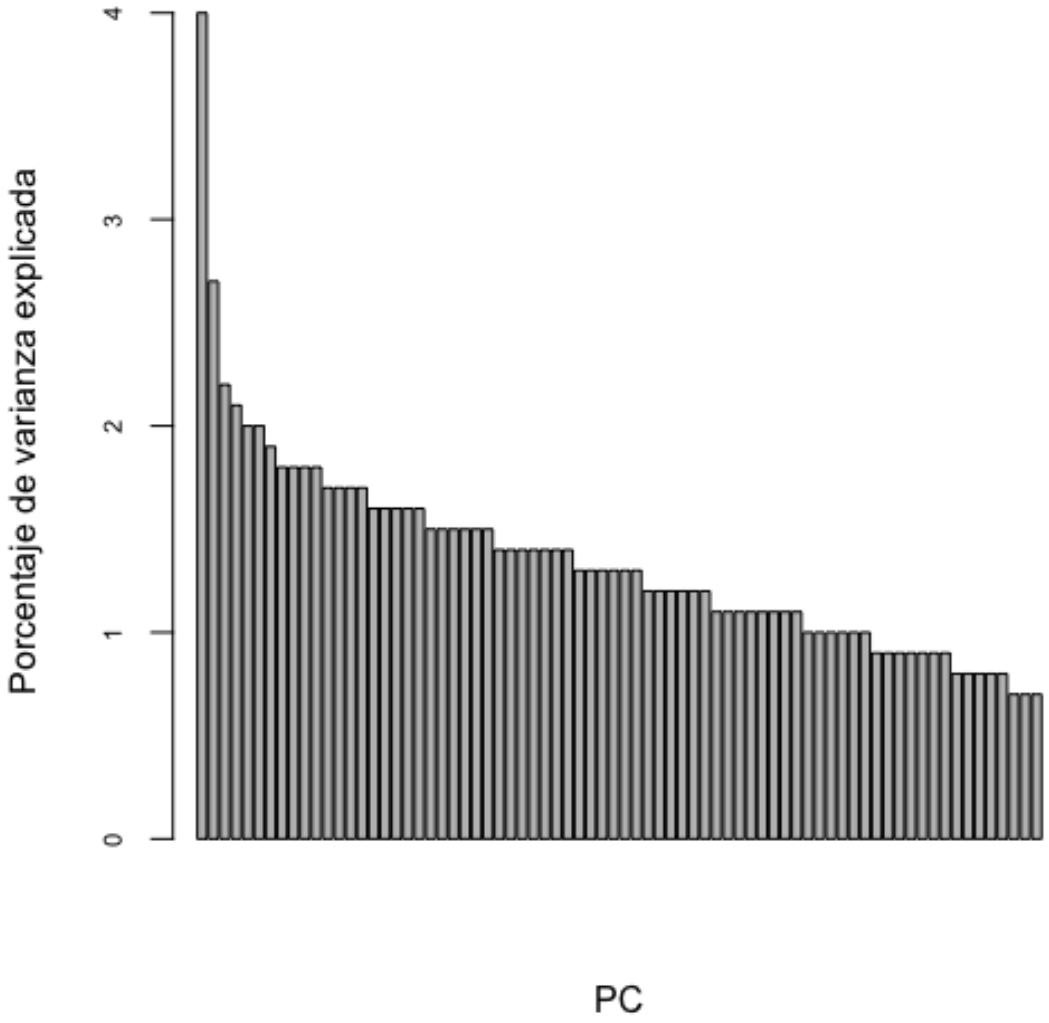

Figura 4.4: Gráfico de sedimentación de los autovalores de cada componente principal (PC) para el subconjunto de *Firmicutes*.

es decir, que la variación total dentro de los mismos sea mínima: a medida que k aumenta, la distorsión promedio disminuye, cada *cluster* tiene menos objetos que lo conforman y los mismas estarán más cerca de sus respectivos centroides.

El método de *elbow* evalúa el WSS total como una función de k mediante la identificación de un punto de quiebre (codo) en la curva, que sugiere el número óptimo de k.

Por otra parte, el análisis de *Silhouette* (S_i) permite determinar qué tan bien un dato fue agrupado en el *cluster*. S_i es calculado de la siguiente manera:

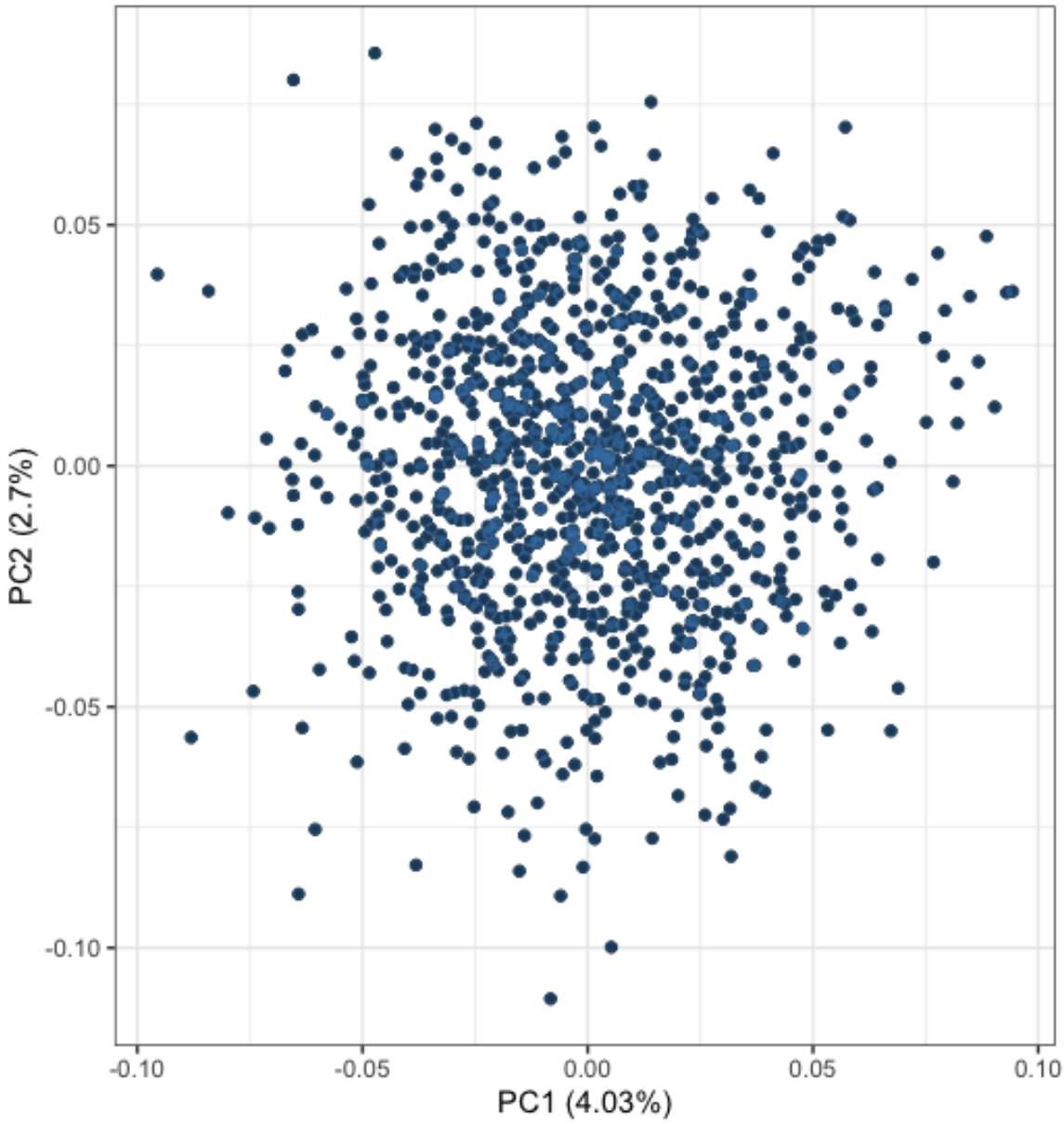

Figura 4.5: *Biplot* representando la posición de cada muestra en términos de PC1 y PC2 para *Firmicutes*.

- Para cada observación i, se calcula la disimilitud a_i entre i y todos las otras observaciones del *cluster* al que pertenece i.

- Para los restantes *clusters* C, a los cuales i no pertenece, se calcula la disimilitud promedio $d(i,\ C)$ de i a todas las observaciones de C. El menor de estos $d(i,\ C)$ se define como $b_i = min_C d(i,\ C)$. El valor de b_i puede verse como la disimilitud entre i y sus *clusters* "vecinos", es decir, el más cercano al cual no pertenece.

- Finalmente, se define S_i de la observación i mediante la fórmula:

$$S_i = (b_i - a_i)/max(a_i, b_i) \tag{4.1}$$

Entre los métodos de evaluación estadística:

- *Gap statistic method*, que se basa en comparar la variación total dentro del *cluster* para diferentes valores de k con respecto a los valores esperados bajo una distribución al azar de referencia, buscando maximizar el estadístico *gap*. Sin embargo, dado que en los conjuntos de datos reales los *clusters* no están totalmente definidos, se busca un balance entre el máximo valor del estadístico y la parsimonia del modelo. Por ello, para determinar el número de *clusters* se identifica el punto a partir del cual el incremento en el estadístico de *gap* comienza a "disminuir".

4.2.1. *Bacteroidetes*

La determinación del número óptimo de k para la implementación de *k-means* indicó un valor de $k = 3$ por el método de *elbow*, $k = 3$ por el coeficiente de *Silhouette* y k = 3 por el *gap method* (Figura 4.6).

Los resultados de *k-means* para $k = 2$ y $k = 3$ mostraron que efectivamente, los datos de *Bacteroidetes* tienen tendencia a distribuírse en dos a tres grupos principales (Figura 4.7).

4.2.2. *Firmicutes*

Para implementar el algoritmo de *k-means* en el subconjunto de *Firmicutes*, se buscó determinar primero el valor óptimo de *k*. El método de *elbow* no indicó ningún valor destacable (Figura 4.8A), el coeficiente de *Silhouette* resaltó $k = 2$ (Figura 4.8B) y finalmente el *gap method* no indicó de manera clara un valor de k (Figura 4.8C). Considerando los tres resultados en conjunto, entonces el subconjunto de *Firmicutes* no es agrupable en *clusters*. Esto queda evidenciado aún más al considerar la distri-

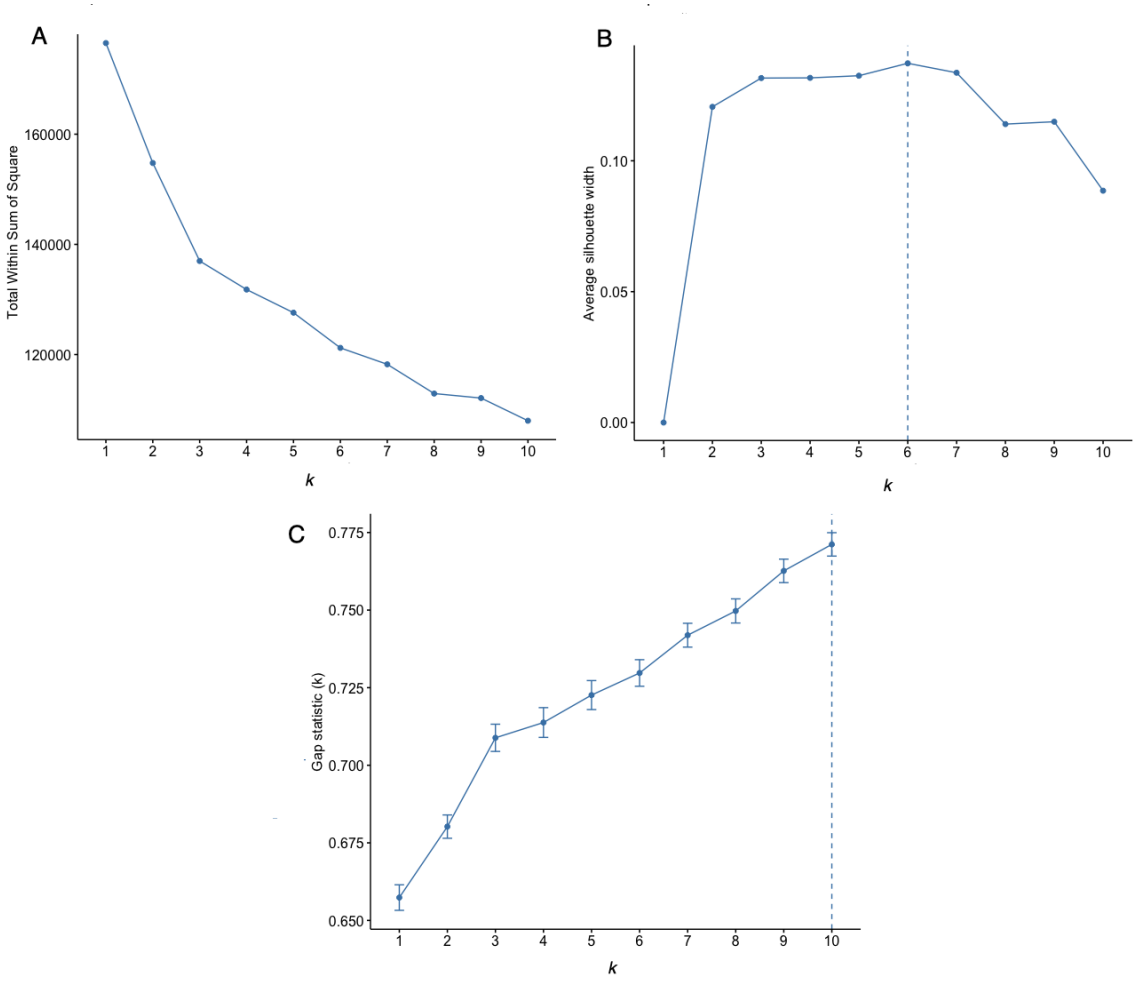

Figura 4.6: Valor óptimo de k para *k-means* mediante (A) el método de *elbow*, (B)Coeficiente de *Silhouette* y (C) *gap method* sobre el subconjunto *Bacteroidetes*.

bución compacta y visiblemente no agrupable de los datos cuando se aplica *k-means* para $k = 1$ (Figura 4.9).

4.3. Conclusiones

Es evidente que los datos del subconjunto de Bacteroidetes se distribuyen de forma natural en 2-3 *clusters*, reflejando la predominancia de los dos géneros más abundantes en este subconjunto, que son *Bacteroides* y *Prevotella*.

En cambio, ningún agrupamiento fue evidente para los datos del subconjunto de *Firmicutes*. Es probable que la distribución compacta, homogénea y sin ningún *cluster* natural refleje el hecho de que la gran mayoría de estas bacterias oscilan entre perfiles de abundancia significativa o por el contrario casi nulas en la mayoría de los individuos,

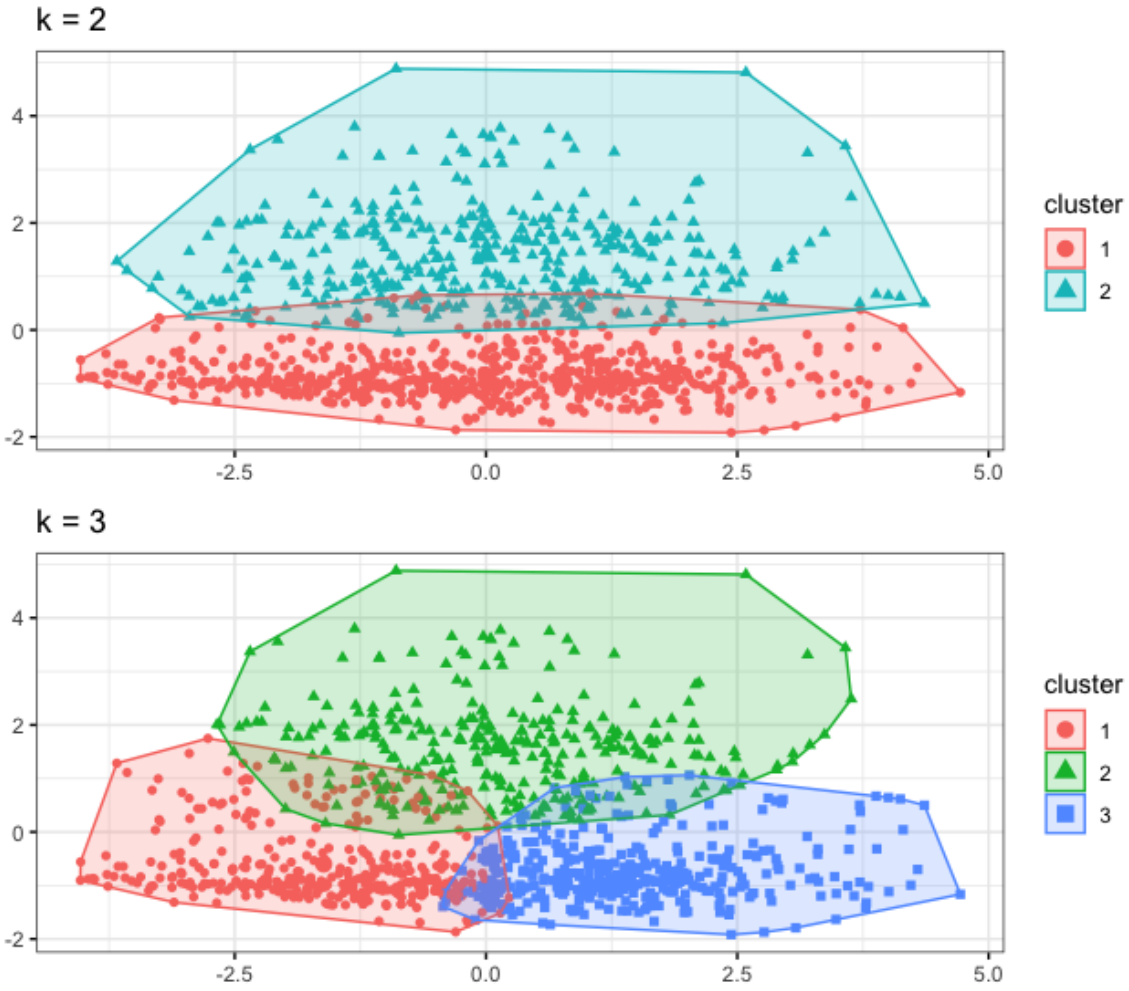

Figura 4.7: Agrupamiento del subconjunto de *Bacteroidetes* mediante *k-means* para diversos valores de *k*.

dificultando la detección de agrupamientos claros (al menos mediante el algoritmo utilizado acá). Se sabe que las abundancias relativas de estas bacterias siguen una distribución bimodal variable [Lahti et al., 2014]. Esta característica podría explicar además la necesidad una cantidad elevada de componentes principales que agrupe un porcentaje significativo de varianza en cada subconjunto.

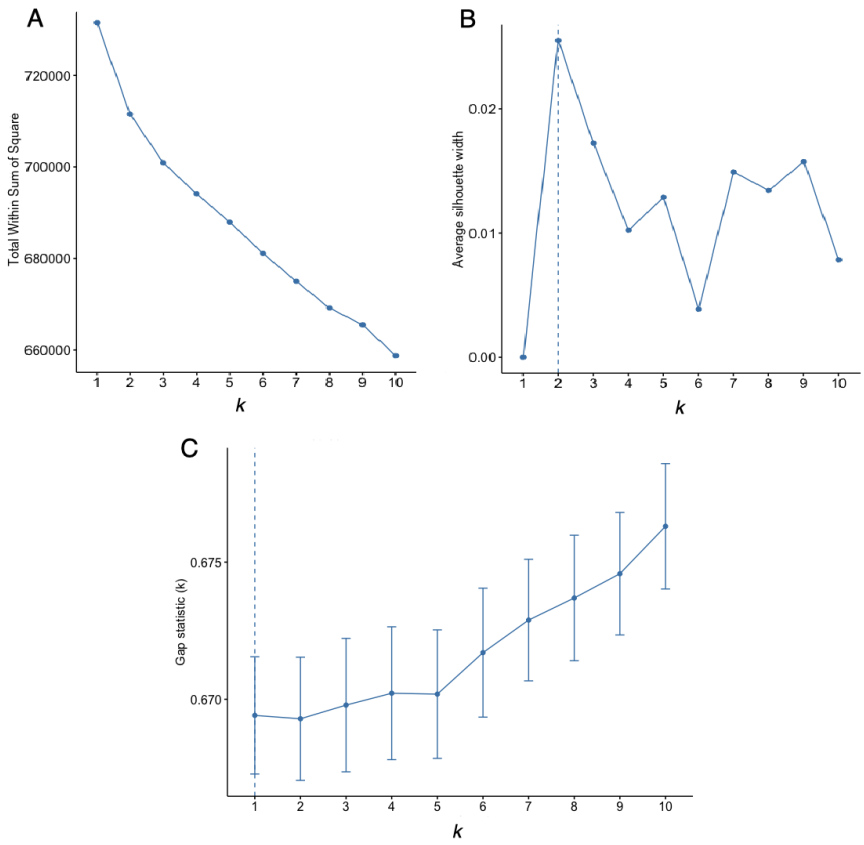

Figura 4.8: Determinación del valor óptimo de k para *k-means* mediante (A) el método de *elbow*, (B) coeficiente de *Silhouette* y (C) *gap method* para el subconjunto de *Firmicutes*.

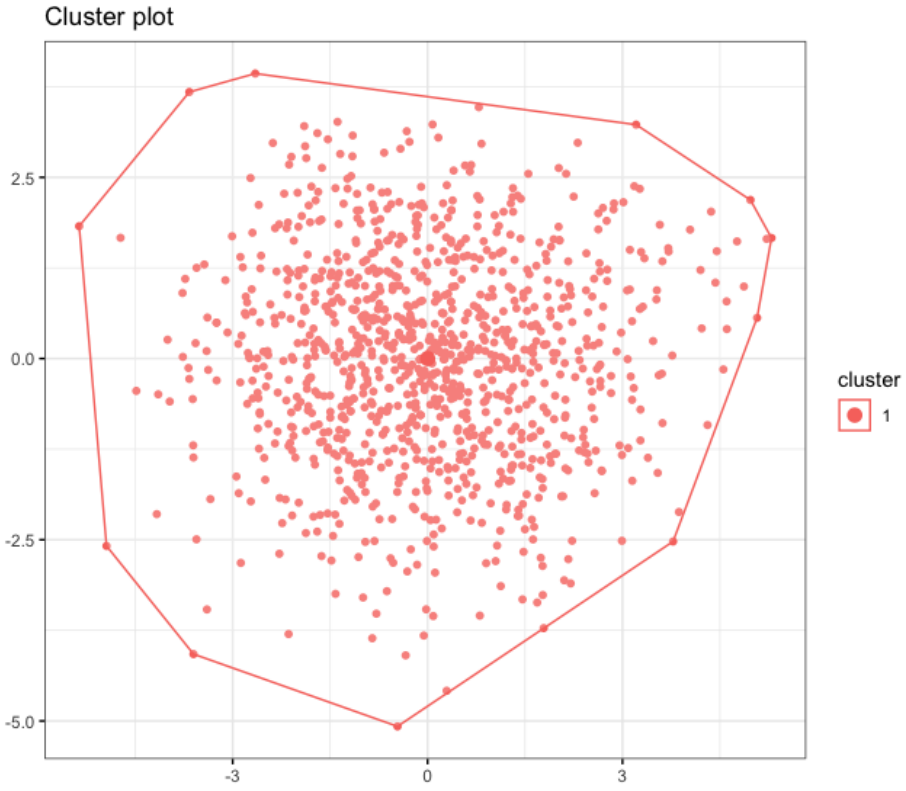

Figura 4.9: Agrupamiento del subconjunto de *Firmicutes* mediante *k-means* para k=1.

Capítulo 5

Mapas Auto-Organizados

Los mapas auto-organizados (SOM, por sus siglas en inglés) son un tipo de red neuronal artificial que usa aprendizaje no supervisado para producir de manera eficaz una representación de datos multidimensionales (vectores) en dos dimensiones. Los SOM, también conocidos como mapas de Kohonen, fueron descriptos por primera vez por Teuvo Kohonen en Finlandia en 1982. Se basan en el aprendizaje competitivo, en el cual las neuronas de salida compiten entre ellas para ser activadas, con el resultado de que sólo una es activada a la vez. Esta neurona activada es la ganadora. Tal competencia puede ser inducida o implementada mediante conexiones inhibitorias laterales entre las neuronas, con el resultado de que las neuronas son forzadas a organizarse (Figura 5.1).

La representación de dos dimensiones (discretizada y de baja dimensionalidad) es un **mapa topográfico**. El concepto de mapa topográfico proviene del área de la neurobiología, el cual establece que las diferentes entradas o *inputs* sensoriales tienen una representación ordenada en su correspondiente área en la corteza cerebral. De esta manera, el mapa tiene dos propiedades importantes:

- Durante la representación, o procesamiento, cada parte de información entrante es mantenida en su propio vecindario.
- Las neuronas que manejan partes de información muy relacionadas se mantienen cercanas de manera que pueden interaccionar mediante conexiones.

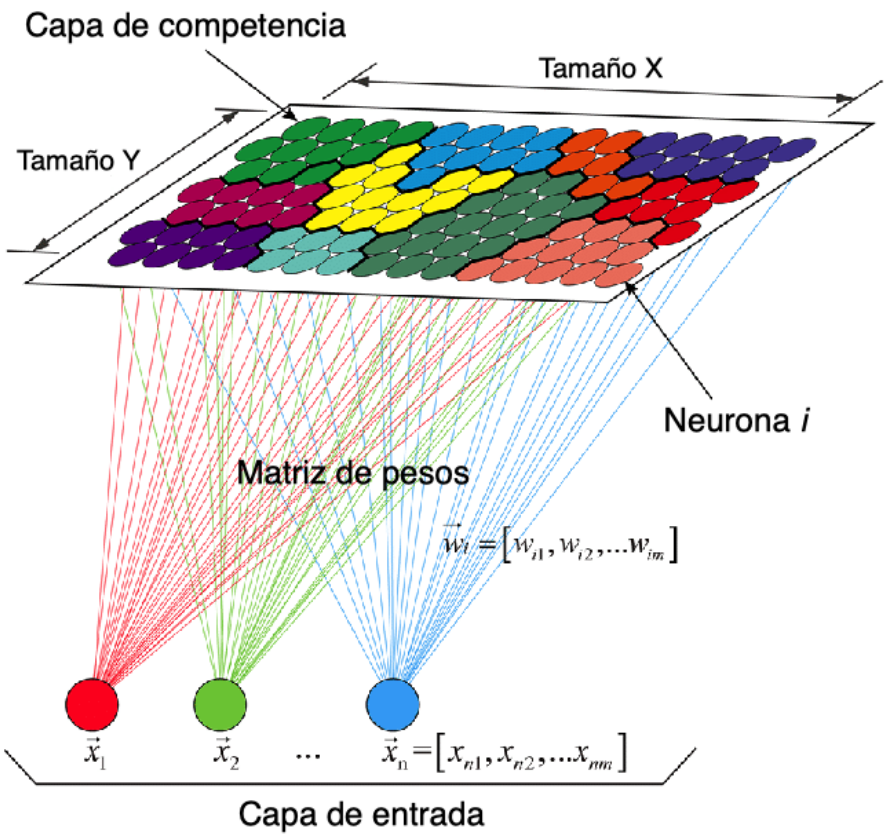

Figura 5.1: Estructura de una red SOM. Los n vectores de la capa de entrada son mapeados en una capa de competencia 2D, representeda por vectores que contienen los pesos. Cada neurona i tiene un vector de peso w de la misma dimensionalidad m que los vectores de entrada(x_n). Adaptado de [Han et al., 2019]

Los SOM tienen una arquitectura de red de dos capas:

- Capa de vectores de entrada, donde ingresa la información de los vectores originales a la red durante el proceso de entrenamiento.

- Capa de nodos de aprendizaje (capa de compentencia), en la cual interesan las relaciones topológicas entre los nodos. Esta capa contendrá finalmente la información acerca de la representación resultante.

Cada dato en la capa de entrada compite por una representación en el SOM. El inicio del algoritmo comienza con la inicialización de los pesos de cada nodo. Luego, se selecciona aleatoriamente un vector de los datos de entrenamiento y se determina en la red de nodos qué pesos son más similares al vector de datos de entrenamiento. El nodo ganador se lo denomina como *Best Matching Unit* (BMU). A partir de esto,

se calcula el vecindario del BMU: más cercanía de un nodo al BMU, sus pesos son más afectados mientras que a más distancia del BMU, aprende menos.

Los pasos a partir de la selección aleatoria de un vector de muestra se repiten durante N iteraciones. De esta manera, el mapa va creciendo y toma diferentes formas: cuadrada, rectangular o hexagonal en el espacio 2D.

La forma de determinar el BMU en cada iteración es mediante el cálculo de la distancia de cada peso al vector de muestra, realizando el cálculo a través de todos los vectores de peso. El peso con la distancia más corta es el ganador. Existen diversos métodos de determinación de la distancia, aunque el más comúnmente usado es la **distancia euclídea**.

Los SOMs son una gran forma de organizar un conjunto de datos multidimensionales en un espacio de dos dimensiones usando una red neuronal. Son usados frecuentemente para tareas de clasificación y *clustering* en diversas áreas como reconocimiento del lenguaje, reconocimiento de imágenes, control de robots y diagnóstico médico, entre otros [Ruan et al., 2011, Gandía-Aguiló et al., 2017].

En el caso de conjuntos de datos multivariados, la visualización de diferentes mapas de calor permite explorar la relación en conjunto entre las diferentes variables dado que todos los mapas están unidos por posición: en cada uno de ellos, un nodo en una determinada ubicación corresponde a la misma unidad en otro mapa. En las siguientes secciones se detallan los resultados de SOM mediante mapas de calor, los cuales son visualizaciones muy eficaces porque muestran la distribución de cada variable a través de la red.

5.1. *Bacteroidetes*

El progreso de aprendizaje de la red para el subconjunto de datos de bacterias pertenecientes a *Bacteroidetes* y variables de metadatos puede ser visualizado mediante una curva que relacione la distancia media al nodo más cercano y el número de iteraciones.

Se observó que a medida que aumentaban las iteraciones, las distancias entre los

nodos hacia el nodo de referencia más cercano iba disminuyendo de manera gradual. Luego de aproximadamente 300 iteraciones, se alcanzó un mínimo estable (Figura 5.2).

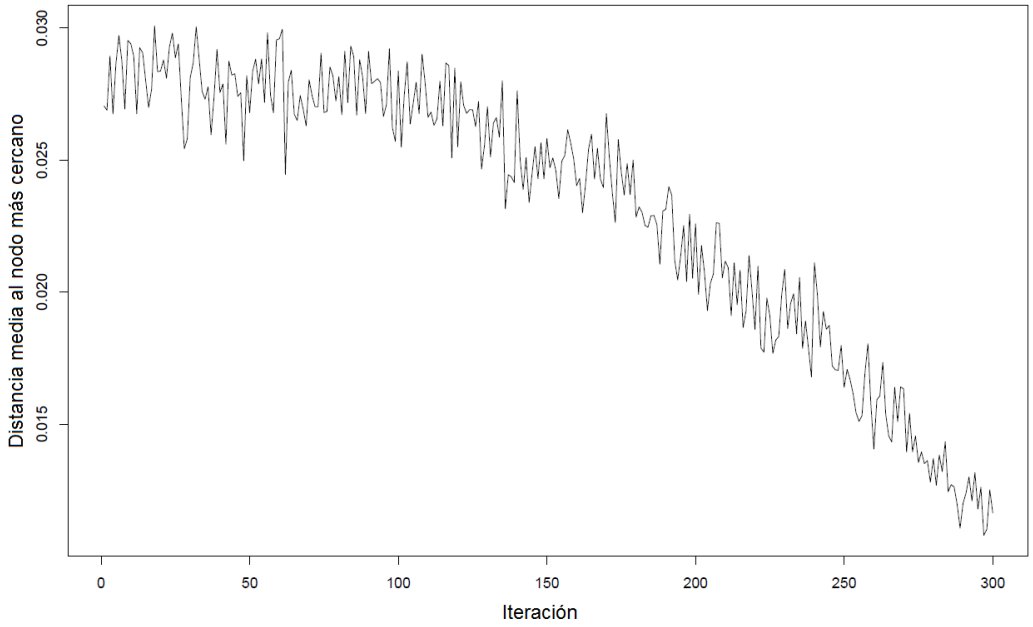

Figura 5.2: Progreso del entrenamiento de la red neuronal SOM en el subconjunto de datos de *Bacteroidetes*.

5.1.1. Metadatos

La distribución de la variable Edad a través del SOM indicó que los nodos correspondientes a rangos de edad más jovenes (alrededor de 20 años) se agruparon en dos regiones del mapa bien definidas y separadas de un pequeño grupo de nodos representando a los individuos de mayor edad ($>$ 65 años). Los individuos del rango de edad media (40-50 años) fueron mapeados en nodos de manera dispersa por todo el mapa (Figura 5.3A, el gradiente de color va desde azul-joven a rojo-mayor).

En cuanto a la distribución de mujeres y hombres en este análisis, se observó que los sexos mapearon en dos grandes grupos marcadamente delimitados (Figura 5.3B).

Tomando solapadamente los mapas que representan las distribuciones de Edad y Sexo se observó que existe solamente una pequeña proporción de individuos jóvenes que se distribuyen tanto en mujeres como hombres. Además, los nodos que agruparon a los individuos de mayor edad correspondieron exclusivamente a hombres. Es decir,

no hay mujeres > 65 años.

La variable Nacionalidad está mayormente representada por individuos de Europa Central. Los mismos mapearon en grupos de nodos bien delimitados, distribuídos a través de todo el mapa. Pocos grupos de nodos aislados mapearon a individuos de Escandinavia. De manera interesante, el mapa identificó un agrupamiento aislado de nodos correspondiente a los individuos de Estados Unidos (Figura 5.3C, nodos color rojo).

Un índice de diversidad ecológica muy utilizado para calcular la diversidad dentro de una población es el índice de diversidad de *Shannon*, el cual calcula el grado de distribución uniforme de las bacterias en una muestra. Se observó que los valores bajos en la escala de diversidad quedaron representados en dos grupos pequeños y aislados de nodos y que, por el contrario, los nodos con valores altos de diversidad mapearon en grupos aislados más grandes y con mayor distribución a través del mapa de la red (Figura 5.3D, el gradiente de color va desde azul-baja diversidad a rojo-alta diversidad).

Los individuos pertenecientes a categorías de BMI bajos (bajo peso-delgados) mapearon de manera aislada en pequeños grupos de nodos. Mientras que los individuos con la categoría más alta de BMI (obesos severos) se distribuyeron principalmente en un conjunto reducido y concentrado de nodos, bien separado de las demás categorías (Figura 5.3E, el gradiente de color va desde azul-bajo a rojo-alto).

Al comparar de manera conjunta los mapas de Diversidad y BMI, se observó que los nodos que representan valores de baja diversidad se superponen con nodos de individuos de BMI alto (es decir, obesos severos). Además, al agregar a esta comparación el mapa de Sexo, se observa que los mismos corresponden al sexo femenino.

Otra característica de la población, desde el análisis de los miembros de *Bacteroidetes* es que los individuos pertenecientes a la población de US son mujeres con valores de diversidad medios y BMI bajo.

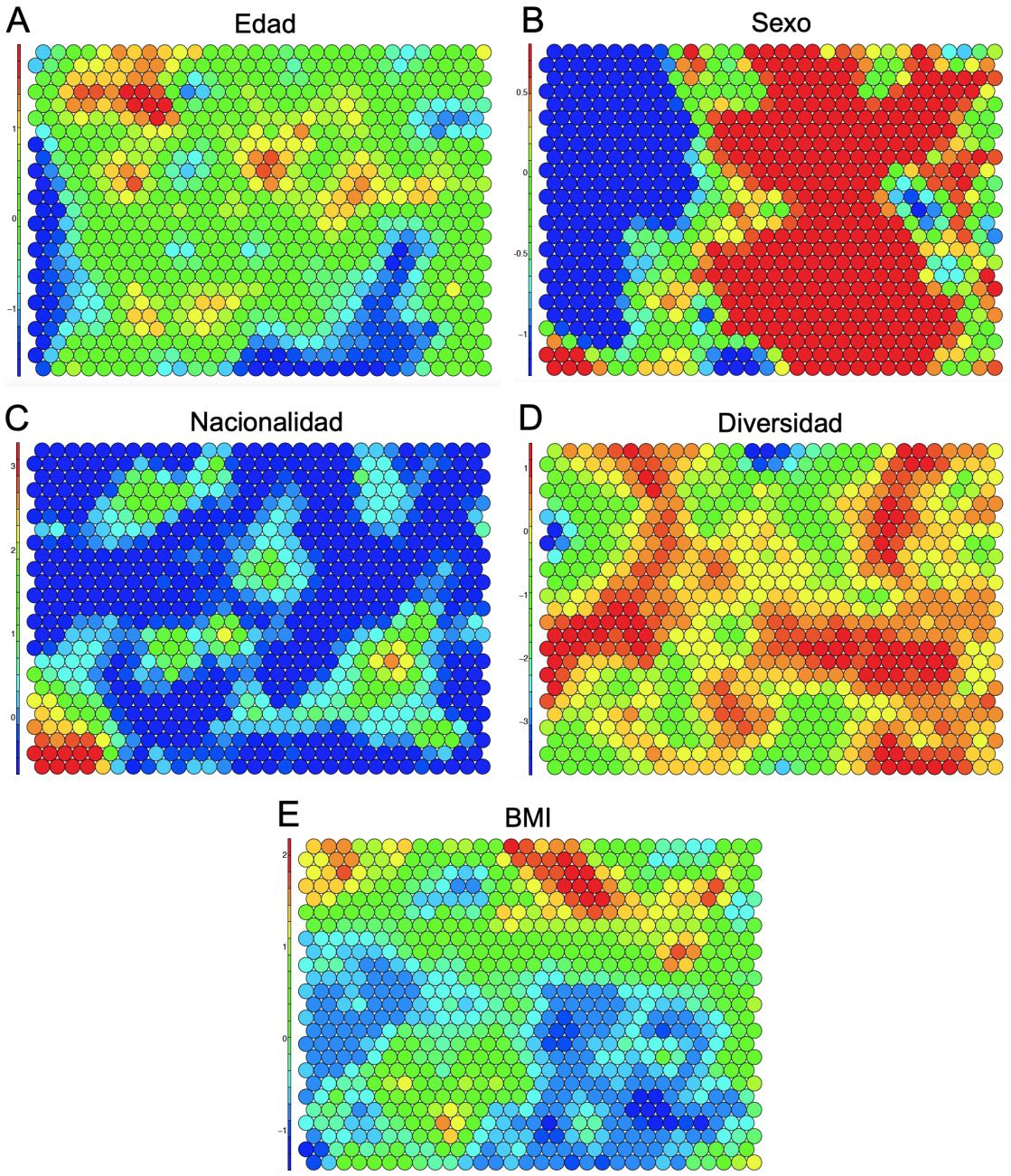

Figura 5.3: Resultados de SOM para las variables de metadata asociados a *Bacteroidetes*. El índice de color de cada mapa está establecido en base a los valores para cada variable. A) Edad, escala ajustada para un rango de 18 a 77 años, azul = más jóvenes, rojo = adultos mayores. B) Sexo, azul = masculino, rojo = femenino. C) Nacionalidad, gradiente de color para azul = Europa Central hasta rojo = US. D) Diversidad, escala ajustada para un rango de 4.7 a 6.3, azul = baja, roja = alta. E) BMI, gradiente de color para azul = delgado hasta rojo = obesos severos.

5.1.2. Abundancias de bacterias

Dentro de la comunidad ecológica de *Bacteroidetes*, los dos géneros más relevantes son *Bacteroides* y *Prevotella*. Las abundancias de ambos están influenciadas por el estilo de vida y dieta: una alta ingesta de grasas y proteínas está asociada con niveles

elevados de *Bacteroides*, mientras que una ingesta alta de fibras está asociada con el niveles altos del segundo género [De Filippo et al., 2010, Wu et al., 2011].

En el subconjunto de datos se identificaron 8 especies pertenecientes al género *Bacteroides*. Estas son: *B. fragilis*, *B. intestinalis*, *B. ovatus*, *B. plebeius*, *B. splachnicus*, *B, stercoris*, *B. uniformis* y *B. vulgatus*.

La distribución de los valores de abundancia relativa para cada especie luego del entrenamiento en la red neuronal mostró un patrón característico: en líneas generales, una gran proporción de nodos mapearon valores de abundancia baja en todo el mapa (color azul). Mientras que para cada especie de bacteria predominó la presencia de pequeños grupos de nodos que concentraron valores altos de abundancia (Figura 5.4, Figura 5.5 y Figura 5.6). Sin embargo, al comparar los mapas entre las especies, no se observó ninguna superposición de los grupos de nodos correspondientes a valores altos de abundancia.

Por otra parte, el género *Prevotella* está representado por 4 especies en el subconjunto de datos. Estas son: *P. melaninogenica*, *P. oralis*, *P. ruminicola* y *P. tannerae*.

Luego del entrenamiento de la red neuronal, se observó que un grupo reducido y delimitado de nodos concentraron los valores más elevados de abundancia relativa de cada especie. De manera similar a lo observado para los miembros de *Bacteroides*, la gran mayoría de los nodos mapearon valores de abundancia baja para las especies de *Prevotella* (Figura 5.5). Al comparar entre las especies, no se observó ninguna superposición de los nodos de máxima abundancia.

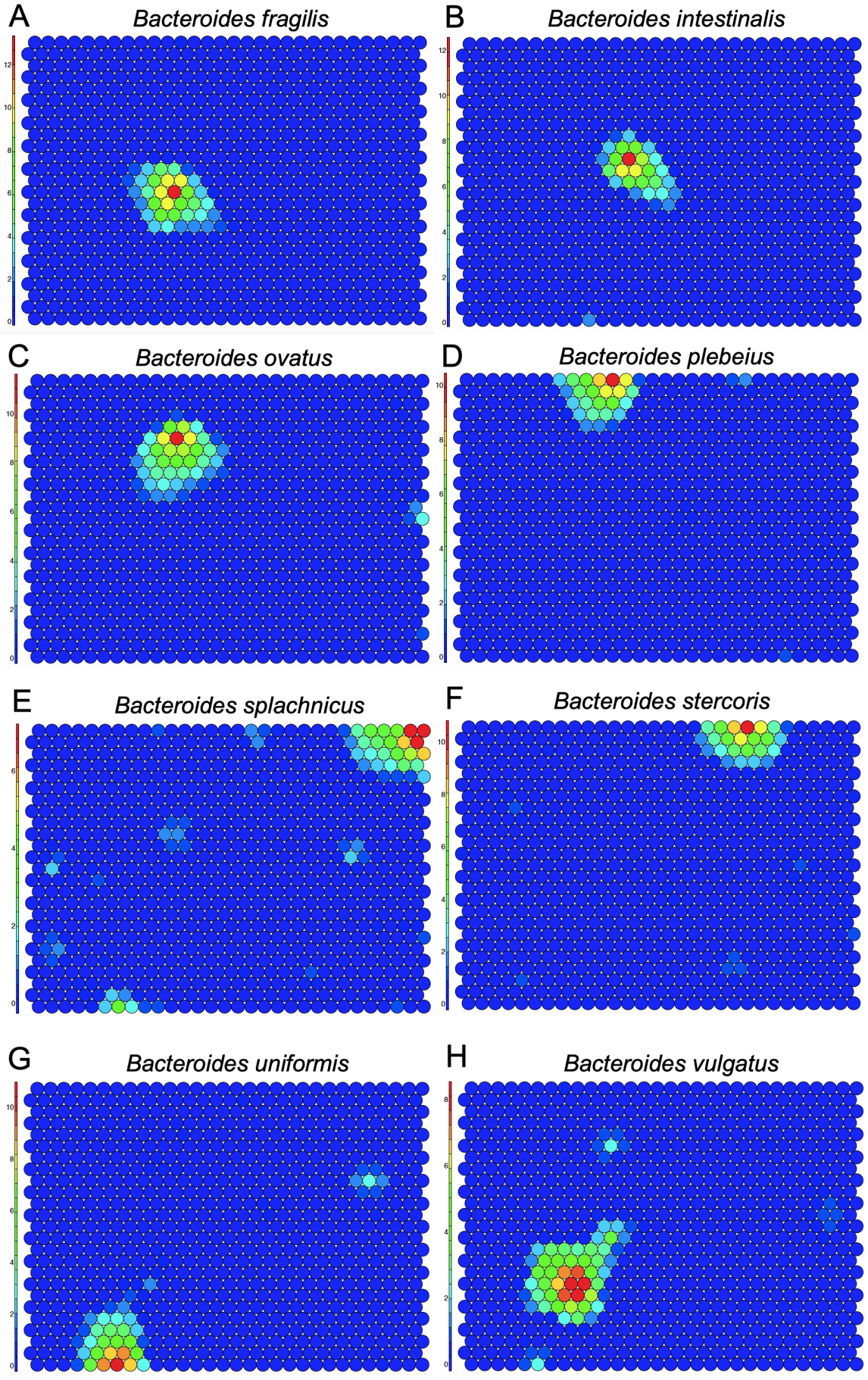

Figura 5.4: Mapas de calor del análisis de SOM para el subconjunto de *Bacteroidetes* enfocado en las especies del género *Bacteroides*. El gradiente de color representa nivel de abundancia. Azul = bajo, rojo = alto.

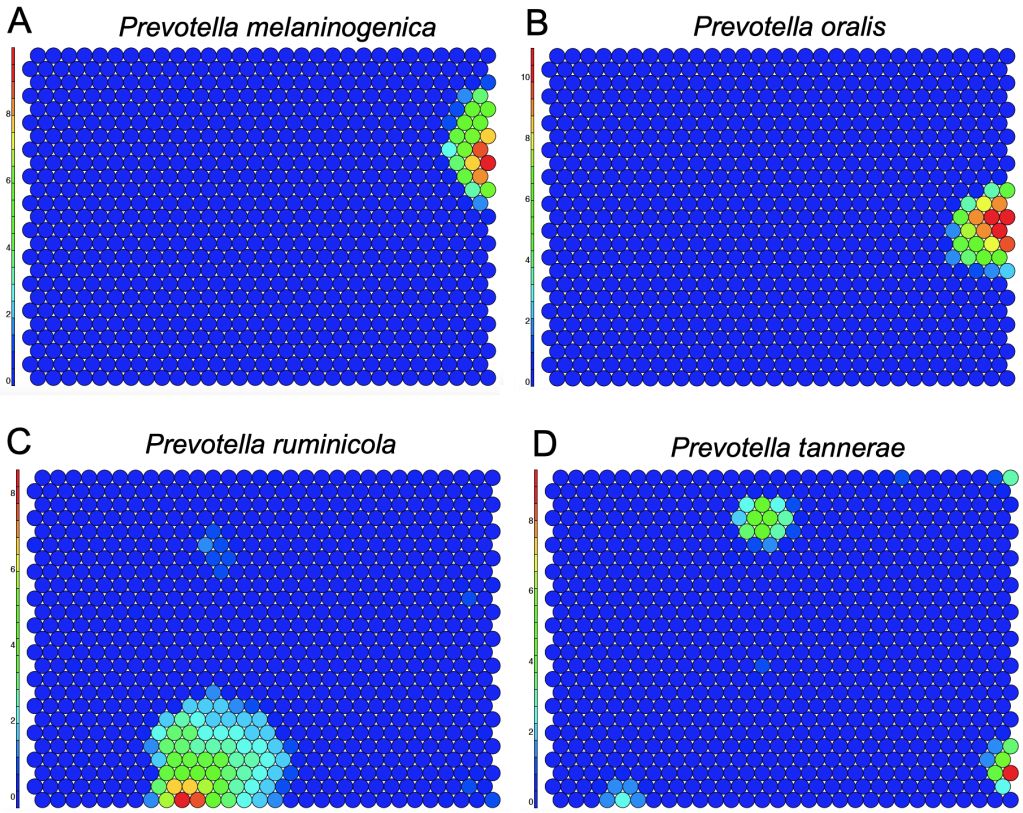

Figura 5.5: Mapas de calor del análisis de SOM para el subconjunto de *Bacteroidetes* enfocado en las especies del género *Prevotella*. El gradiente de color representa nivel de abundancia. Azul = bajo, rojo = alto.

En cuanto a los restantes géneros de *Bacteroidetes* identificados en el subconjunto de datos, tales como *Allistipes*, *Tannerella* y *Parabacteroides distasonis*, se observó que los niveles de máxima abundancia relativa se distribuyeron en grupos aislados y concentrados de nodos en el mapa (Figura 5.6). De manera similar a los resultados mencionados anteriormente, no se observó superposición entre los nodos de máxima abundancia entre las especies.

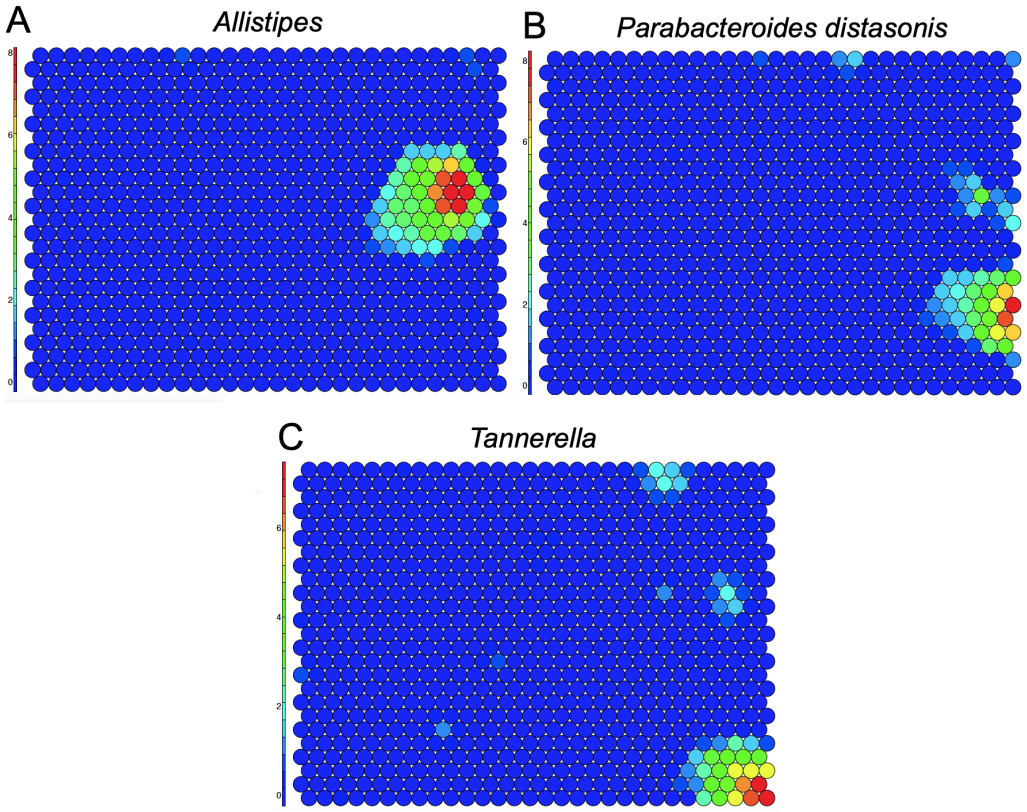

Figura 5.6: Mapas de calor del análisis de SOM para el subconjunto de *Bacteroidetes* enfocado en especies de los géneros *Allistipes*, *Tannerella* y *Parabacteroides*. El gradiente de color representa nivel de abundancia. Azul = bajo, rojo = alto.

Si bien la red neuronal SOM genera una proyección 2D de los datos que ayuda a determinar visualmente características y probables agrupamientos, es posible complementar el enfoque con el uso de un algoritmo de *clustering* aplicado a los vectores prototipo encontrados por la SOM y así dividir los datos del mapa 2D en *clusters*. En el caso más simple, se puede considerar a cada neurona como un único *cluster* o se pueden agrupar varias neuronas en un *cluster*. De esta manera, el análisis de los mapas de calor 2D para cada variable en estudio se completa mediante la identicación manual de los *clusters*, y en conjunto, se les da sentido a los resultados acerca de las diferentes áreas en el mapa.

Se aplicó un algoritmo de agrupamiento jerárquico sobre los datos de tal manera que los *clusters* quedaron implementados sobre los mapas de SOM. Inicialmente se probaron varias condiciones para el número de *clusters*, llegando a la condición n=16 grupos.

Los resultados combinados de SOM y *clustering* son mejor visualizados mediante *boxplots* para así comparar las características de cada *cluster*: Son 16 *clusters*, denominados de A a P, y cada uno representa atributos particulares de un grupo de nodos (Fig. 5.7 a Fig. 5.10).

Hay *clusters* que tienen características más definidas que otros y, por lo tanto, es más directa su descripción.

- Es posible identificar niveles de abundancia de las especies de *Bacteroides* en los *clusters* B, F, I, K, M, N, O y P (Fig. 5.8).
- Las especies de *Prevotella* están agrupadas en los nodos C, E, H, L, M y N (Fig. 5.9).
- Mientras que los tres restantes géneros de *Allistipes Parabacteroides* y *Tannerella* abundan en los *clusters* D, G, J y P (Fig. 5.10).

Cluster D:

- Solamente se identificó una abundancia detectable para una única especie del género *Tannerella*.
- Comprende a individuos de 50 años.
- Existe una tendencia a que sean mujeres.
- La nacionalidad es Europa Central.
- La diversidad biológica es de las más altas entre los clusters.
- El índice de BMI corresponde a individuos con sobrepeso.

Cluster I:

- Dos especies se identifican en niveles de abundancia detectables: *Bacteroides fragilis* y *Bacteroides vulgatus*.
- Los individuos tienen 50 años.

- No queda definido la predominancia de sexo en este grupo.
- La diversidad biológica es intermedia.
- El índice de BMI corresponde a individuos delgados.

Cluster M:

- Solamente dos especies están presentes en niveles detectables: *Bacteroides ovatus* y *Bacteroides vulgatus*.
- Los individuos de este *cluster* tienen alrededor de 50 años.
- No hay una definición clara de sexo, pero existe mayor probabilidad de que sean mujeres.
- La diversidad biológica para este grupo es de las más bajas entre los *clusters*.
- El índice de BMI corresponde a individuos con sobrepeso.

Cluster P:

- En este grupo se identificaron niveles de abundancia detectables para tres especies: *Bacteroides plebeius*, *Bacteroides stercoris* y *Tannerella*.
- Los individuos tienen alrededor de 45 años.
- Existe mayor tendencia al género femenino.
- Pertenecen a Escandinavia.
- La diversidad biológica para este grupo es una de las más bajas, similar a la del *cluster* M.
- El índice de BMI corresponde a individuos obesos.

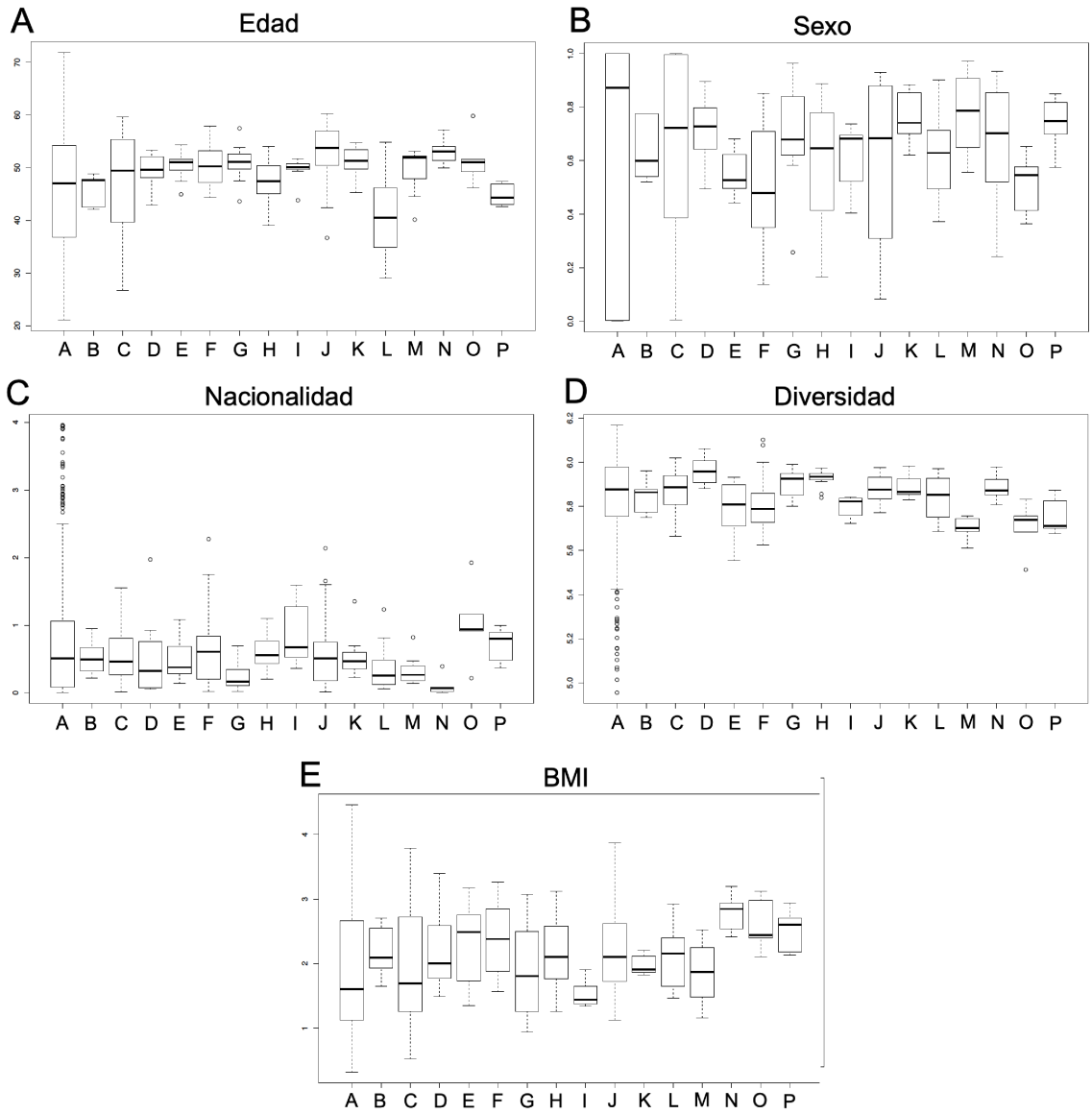

Figura 5.7: Combinación de SOM y *clustering* jerárquico sobre el subconjunto de *Bacteroidetes*. A) Edad. B) Sexo, 1 = femenino, 0 = masculino. C) Nacionalidad, 0 = Europa Central, 1 = Escandinavia, 2 = Europa del Sur, 3 = Reino Unido/Irlanda, 4 = Estados Unidos. D) Diversidad. E) BMI, 0 = bajo peso, 1 = delgado, 2 = sobrepeso, 3 = obeso, 4 = obeso severo, 5 = mórbido obeso.

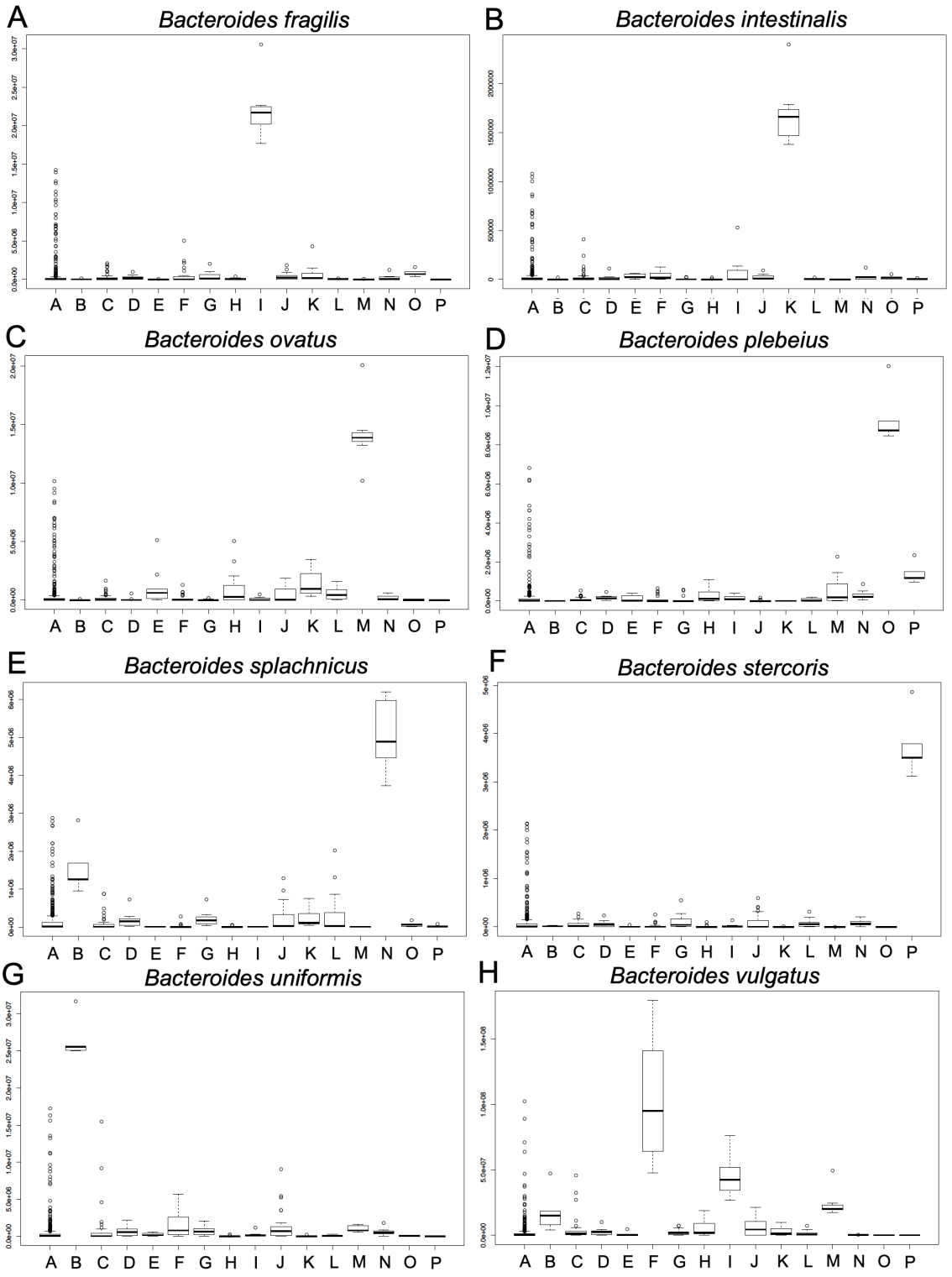

Figura 5.8: Comparación de *clusters* para las especies *Bacteroides fragilis* y *Bacteroides vulgatus* luego del agrupamiento y SOM.

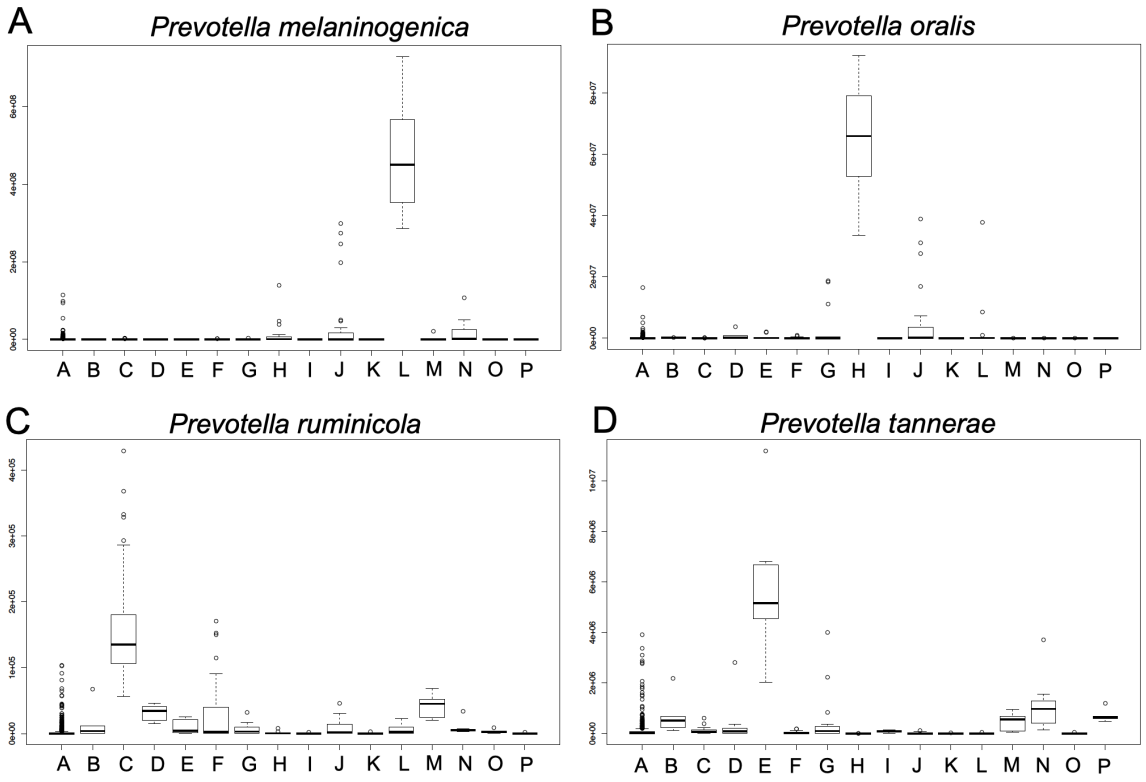

Figura 5.9: Comparación de *clusters* teniendo en cuenta la diversidad y el grado de BMI luego del agrupamiento y SOM. Las categorías de BMI corresponden a 0 = bajo peso, 1 = delgado, 2 = sobrepeso, 3 = obeso, 4 = severo obeso, 5 = mórbido obeso.

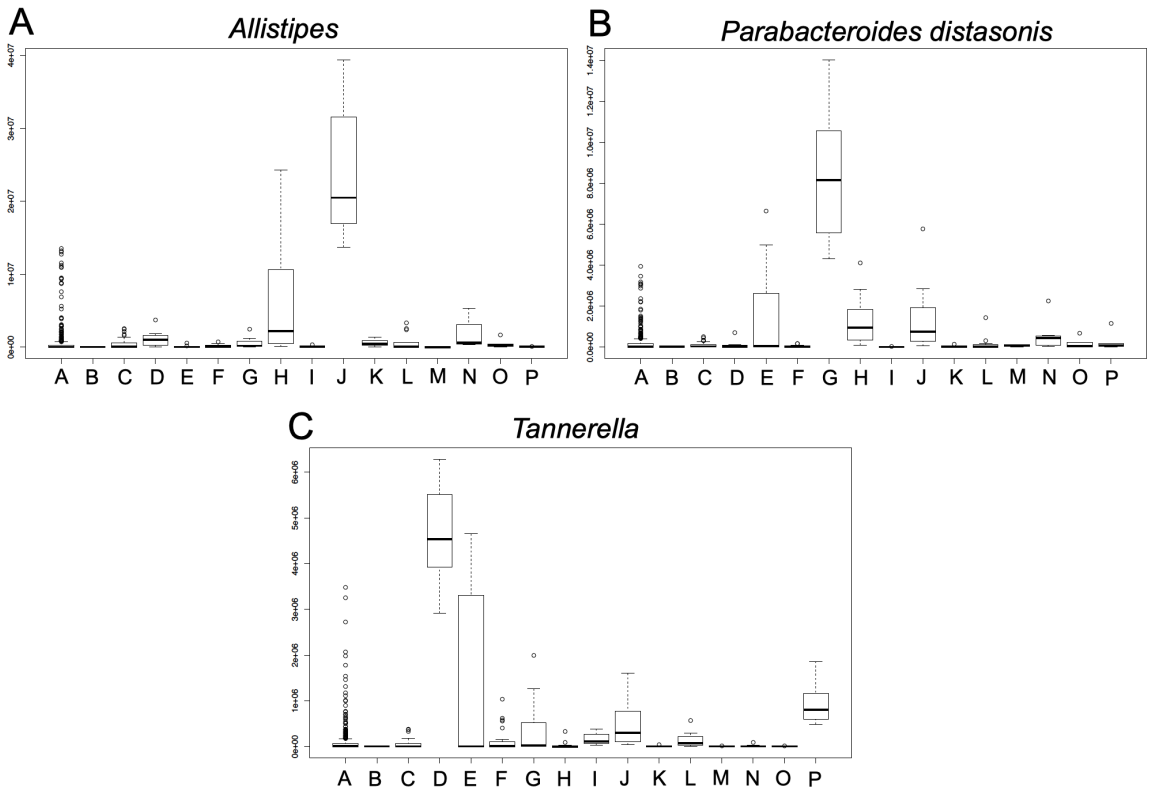

Figura 5.10: Comparación de *clusters*.

5.2. *Firmicutes*

De manera similar a lo observado con el subconjunto de *Bacteroidetes*, a medida que aumentaban las iteraciones, la curva de entrenamiento del SOM disminuyó de manera gradual, indicando el aprendizaje de la red. Luego de aproximadamente 300 iteraciones, se alcanzó un mínimo estable (Figura 5.11)

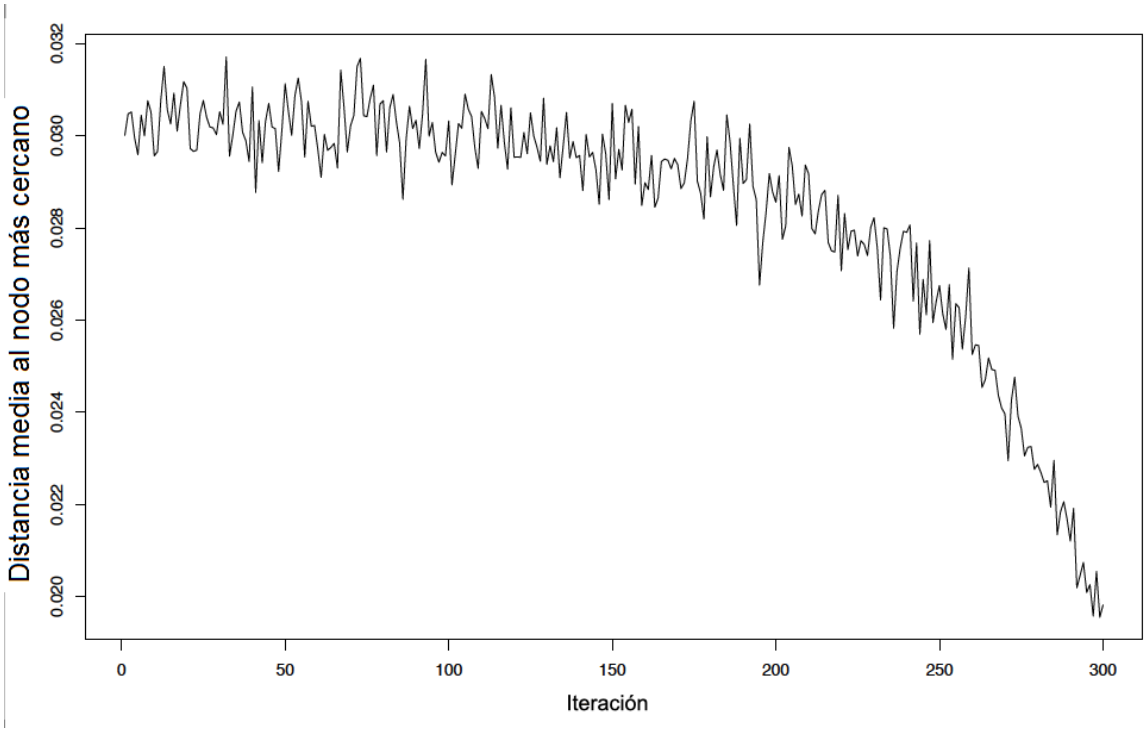

Figura 5.11: Progreso del entrenamiento de la red neuronal SOM en los datos de *Firmicutes*.

5.2.1. Metadatos

Cuando las variables de metadatos son entrenadas en la red SOM asociadas a *Firmicutes* resultaron en distribuciones más dispersas de grupos de nodos en comparación a lo observado con *Bacteroidetes*.

La variable Edad mostró en su mayoría una distribucion homogénea de valores de edad media en todo el mapa (40-50 años), mientras que se observó unos muy pocos nodos aislados correspondientes a los individuos más jóvenes (alrededor de 20 años) o a los individuos $>$ a 65 años (Figura 5.12A).

En lo referente a la distribución de sexos, no se observó una separación muy evidente. Solamente algunos grupos de nodos aislados mapearon la probabilidad de ser hombres o mujeres (Figura 5.12B).

Dado que la variable Nacionalidad está mayormente representada por individuos de Europa Central, se observó que los mismos mapearon en grupos de nodos bien delimitados, distribuídos a través de todo el mapa. Pocos grupos de nodos aislados mapearon a individuos de Escandinavia. De manera interesante, el mapa identificó un

agrupamiento aislado de nodos correspondiente a los individuos de US (Figura 5.12C).

La diversidad biológica para *Firmicutes* mostró grupos de nodos correspondientes en su mayoría a valores de diversidad altos, observándose solamente tres nodos que mapearon para valores de diversidad bajo (Figura 5.12D).

En cuanto a la distribución del índice de BMI, se observó una predominancia de valores bajos de BMI, correspondientes a individuos delgados y/o de bajo peso. Solamente un grupo muy reducido de nodos agruparon a los valores más altos de BMI, correspondiente a la categoría obesos severos (Figura 5.12E).

Al comparar de manera conjunta los mapas de diversidad y BMI, se observó una superposición de ciertos grupos de nodos que representan valores de máxima diversidad con nodos que agrupan categorías de BMI bajo.

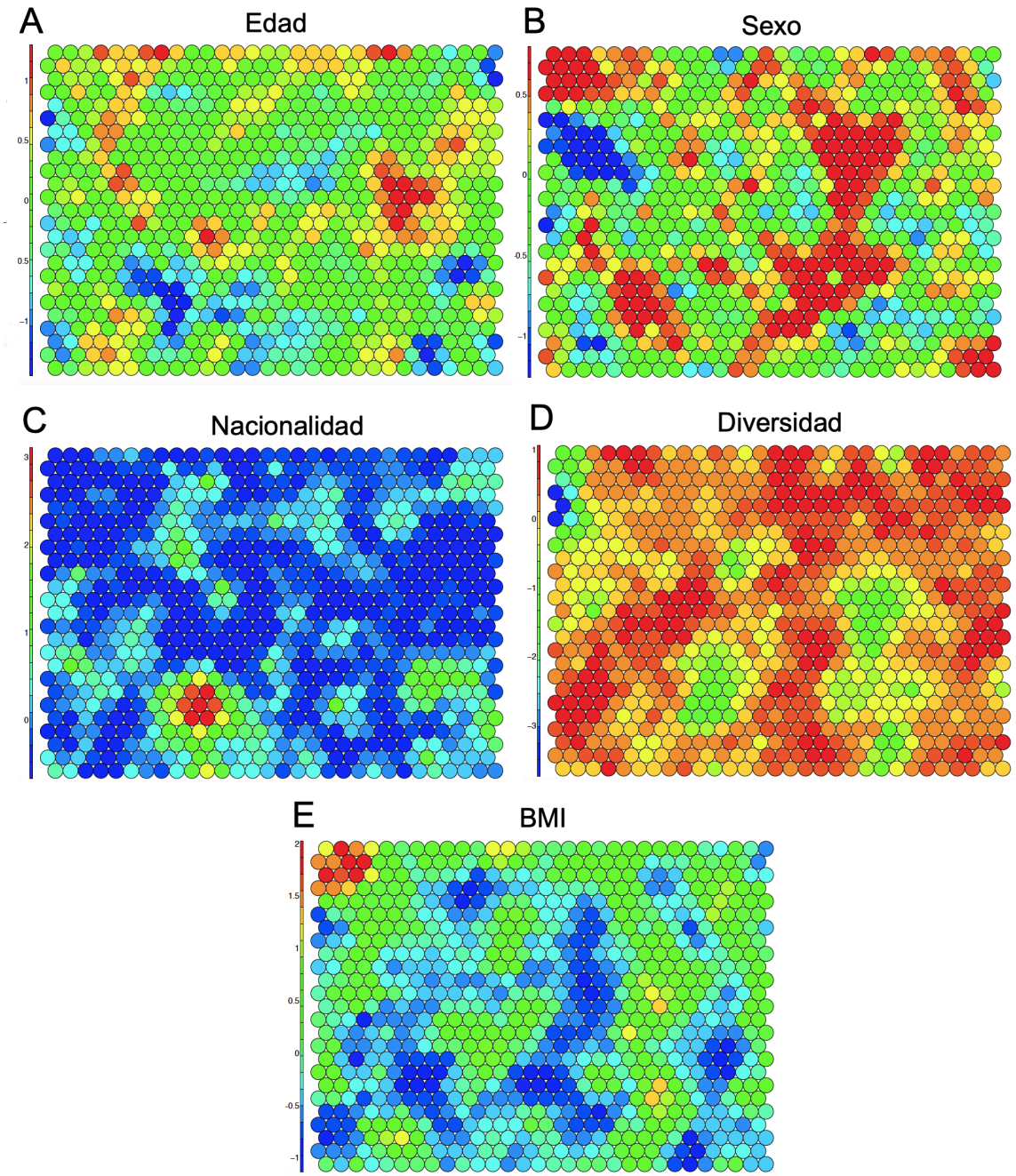

Figura 5.12: Resultados de SOM para las variables de metadata correspondientes a la red de *Firmicutes*. El índice de color de cada mapa está establecido en base a los valores para cada variable. A) Edad, escala ajustada para un rango de 18 a 77 años, azul = más jóvenes, rojo = adultos mayores. B) Sexo, azul = masculino, rojo = femenino. C) Nacionalidad, gradiente de color para azul = Europa Central hasta rojo = US. D) Diversidad, escala ajustada para un rango de 4.7 a 6.3, azul = baja, roja = alta. E) BMI, gradiente de color para azul = delgado hasta rojo = obesos severos.

5.2.2. Abundancias de bacterias

El *phylum Firmicutes* está conformado por alrededor de 250 géneros diferentes de bacterias, tales como *Lactobacillus*, *Bacillus*, *Clostridium*, *Enterococcus* y *Ruminicoc-*

cus, entre otros. En el subconjunto de datos analizado en esta parte de la tesis se identificaron 74 miembros pertenecientes a este *phylym*, cada uno de los cuales generó un mapa de calor con los resultados del entrenamiento en la SOM. La interpretación visual de todos los mapas en búsqueda de una superposición biológica interesante resulta difícil de realizar a simple vista.

Es por ello que, teniendo en cuenta que el subconjunto de datos contiene información de las diferentes categorías de masa corporal de los individuos y valores de diversidad del microbioma, se consideró interesante enfocar el análisis de los mapas en una determinada cantidad de miembros específicos de *Firmicutes* que previamente han sido asociados la obesidad [Koliada et al., 2017] [Crovesy et al., 2020]. Sin embargo, los mapas de las bacterias de *Firmicutes* no mostradas en esta parte están disponibles en la sección de Anexos. Las bacterias seleccionadas para mostrar los mapas luego del entrenamiento en la SOM son miembros de los géneros *Clostridium*, *Eubacterium*, *Lactobacillus* y *Ruminicoccus*.

Se seleccionaron también los mapas de otros miembros de *Firmicutes*, como *Aneurinibacillus* y *Peptostreptococcus micros* dado que se observó una superposición de los niveles de abundancia de estos con valores extremos de BMI correspondientes a individuos delgados u obesos.

Los resultados del análisis de SOM para estos miembros indicaron que los valores bajos de abundancia relativa se distribuyeron en la mayoría de los nodos del mapa, mientras que solamente pequeños grupos aislados de nodos concentraron los valores altos de abundancia para cada bacteria, sin existir superposición entre ellos.

Es interesante indicar que solamente un nodo correspondiente a individuos obesos coincidió con la abundancia elevada de una única especie de *Firmicutes*, tal como se muestra en la Figura 5.13G para *Peptostreptococcus micros*. No se observaron otras superposiciones de nodos correspondientes a BMI elevados con ninguna otra de las bacterias seleccionadas. Entonces, la presencia de solamente una bacteria quedó claramente asociada a personas obesas.

Por otra parte, se observó una superposición variada entre los nodos que agruparon

individuos delgados y de bajo peso (índices de BMI bajos) con los nodos que agruparon niveles de abundancia elevada para las siguientes bacterias: *Aneurinibacillus*, *Clostridium sensu stricto*, *Clostridium felsineum*, *Clostridium thermocellum*, *Eubacterium ventriosum*, *Lactobacillus plantarum*, *Ruminococcus bromii* (Figura 5.13A-F, H). Es decir, la presencia de las especies mencionadas aparece asociada a personas delgadas.

5.3. Conclusiones

Es posible observar que ambos subconjuntos de datos funcionan de manera diferente para algunas variables de metadatos (Figura 5.14):

La discriminación de individuos en cuanto a Edad y Sexo fue más efectiva usando los datos de *Bacteroidetes*.

La identificación de individuos de acuerdo a las diferentes categorías de BMI fue más efectiva usando los datos de *Firmicutes*.

Los nodos que agruparon individuos delgados estuvieron principalmente asociados a abundancias altas de ciertos de *Firmicutes*.

Los nodos que agruparon individuos con sobrepeso u obesos estuvieron asociados a abundancias altas de varias especies de *Bacteroidetes*.

De acuerdo a los resultados observados es relevante destacar el uso de SOM en el estudio de un conjunto de datos multivariado, que combina metadatos y expresión génica.

El análisis de SOM es único en el sentido de que combina dos aspectos, el de la proyección y el agrupamiento. Mientras que el agrupamiento brinda información sobre patrones de expresión similares (en este caso, la expresión del gen 16S rRNA relacionado directamente con la abundancia de un determinado tipo de bacteria), la proyección de datos a un plano sirve como un método de visualización y permite estudiar las relaciones entre los datos.

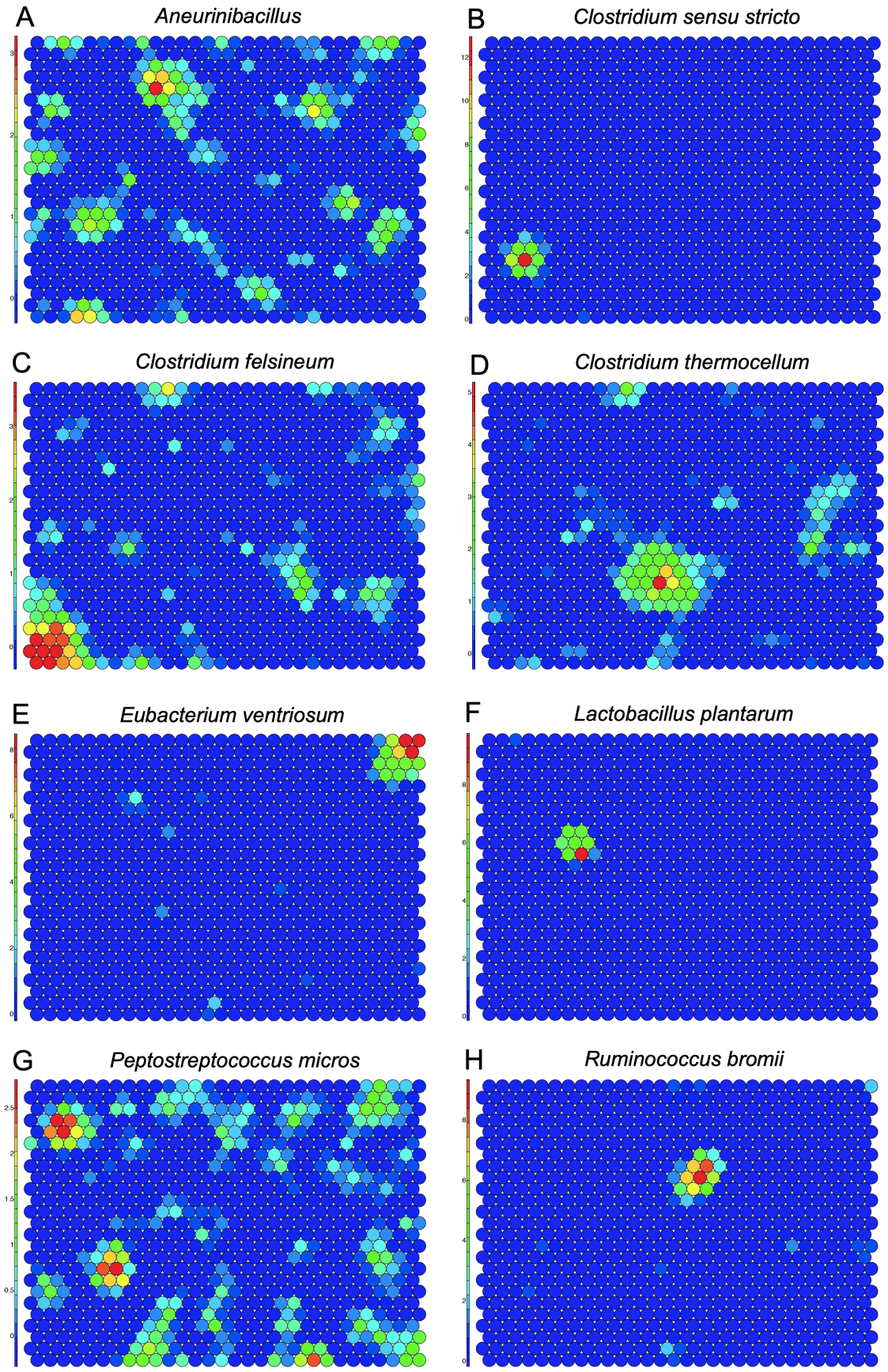

Figura 5.13: Mapas de calor del análisis de SOM para el subconjunto de *Firmicutes* enfocado en bacterias del género *Aneurinibacillus* (A), *Clostridium* (B-D), *Eubacterium* (E), *Lactobacillus* (F), *Peptostreptococcus* (G) y *Ruminicoccus* (H). El gradiente de color representa el nivel de abundancia. Azul = bajo, rojo = alto.

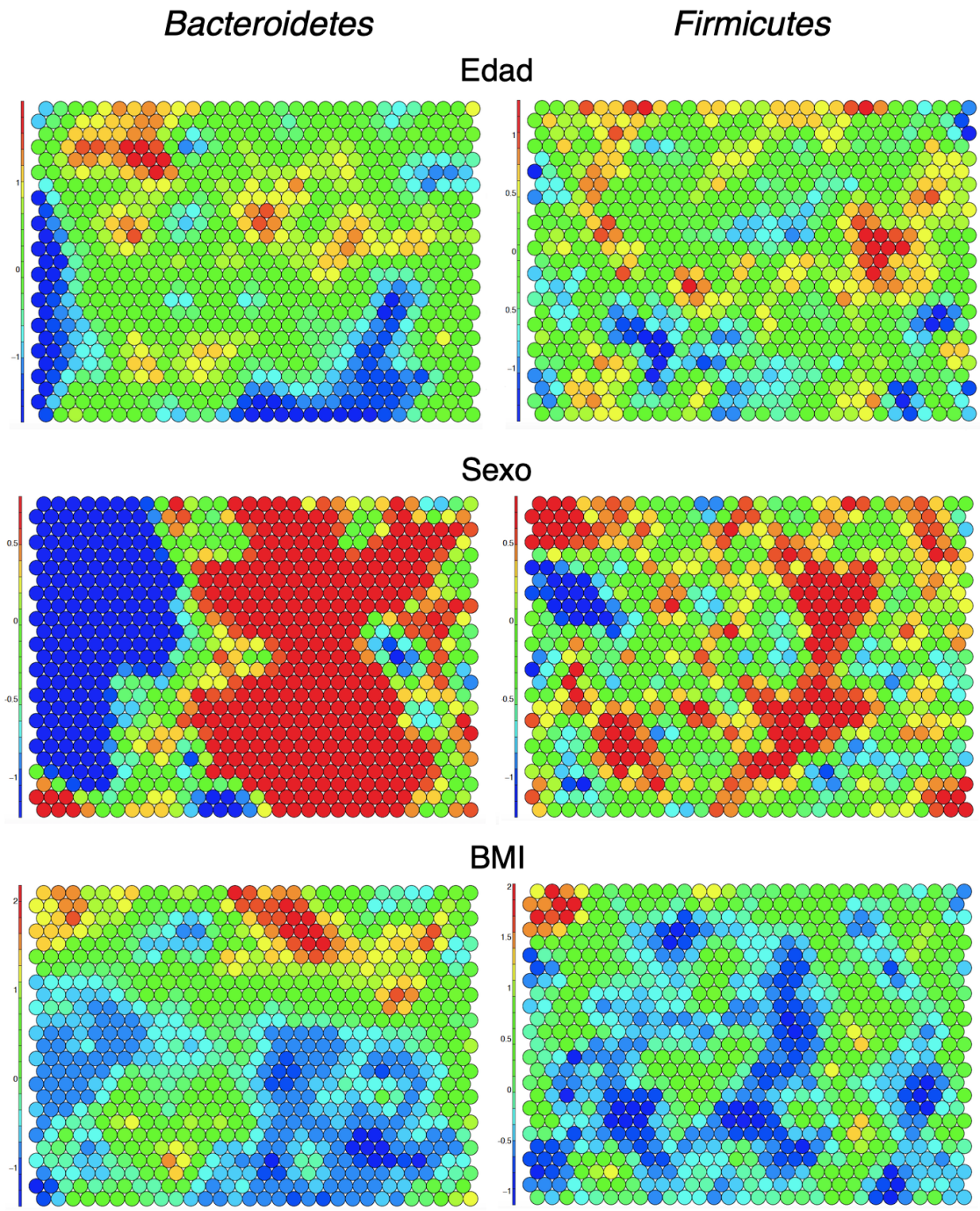

Figura 5.14: Comparación de las variables de metadatos discriminadas de manera diferente por los subconjuntos *Bacteroidetes* y *Firmicutes*.

Capítulo 6

Random Forest

Random forest (RF) es un algoritmo de aprendizaje supervisado muy utilizado en ML que produce, incluso sin ajuste de hiperparámetros, resultados robustos y de impacto en problemas de toma de decisiones, tanto para clasificación como para regresión. RF está basado en árboles de decisión los cuales tienen una estructura de grafo abierto, similar a la estructura de un árbol: cada nodo interno representa un atributo, cada rama representa una decisión y cada nodo hoja terminal representa un resultado categórico o contínuo (es decir, la decisión tomada luego de calcular todos los atributos). El camino desde el nodo raíz hasta el nodo hoja representan reglas de clasificación.

Los algoritmos inductores de árboles de decisión son considerados los métodos de aprendizaje supervisado más usados. Todos comparten el objetivo de encontrar el árbol óptimo, de tal manera que cada partición binaria va dividiendo el árbol de manera recursiva en nodos terminales cada vez más homogéneos en relación a una variable *target*, minimizando el error.

Un esquema típico para un algoritmo de inducción *top-down* es el que se indica en la Figura 6.1. En cada iteración, el algoritmo considera la partición del conjunto de entrenamiento usando la salida de una función discreta de los atributos de inicio. Luego de la selección de la partición apropiada, cada nodo continúa subdividiendo el conjunto de entrenamiento en subconjuntos más pequeños, hasta que se satisface algún criterio previamente establecido.

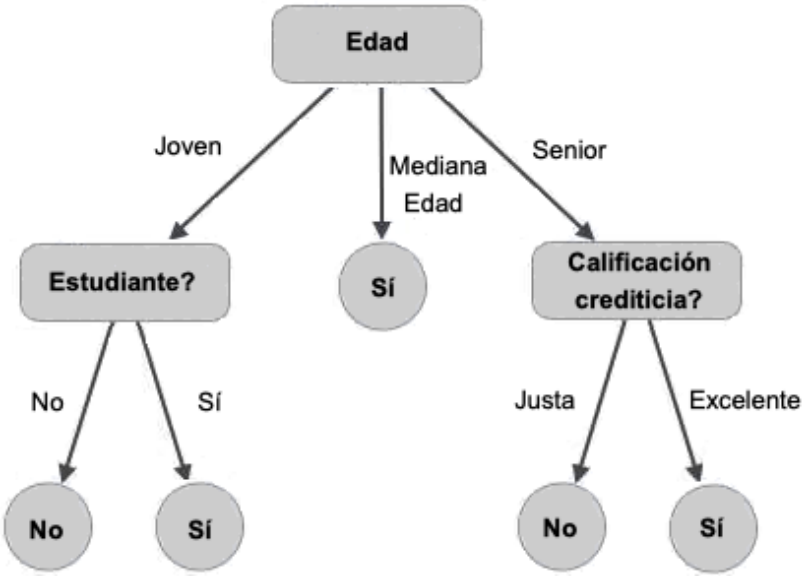

Figura 6.1: Ejemplo de árbol de decisión.

.

Sin embargo, un árbol de decisión simple con poca profundidad resulta en un modelo débil al momento de analizar conjuntos de datos complejos y grandes. Por ello, para obtener mejoras en la eficiencia de predicción se recurre a los ensambles, que son una combinación de varios modelos de *machine learning* en un modelo fuerte. Los métodos de ensamble son efectivos ya que la combinación de resultados de múltiples modelos reduce el error de varianza.

Un procedimiento de ensamble es *bagging* (o *bootstrap aggregation*) que toma al azar varios subconjuntos de datos a partir de los datos de entrenamiento. Así, cada colección de subconjuntos es usada para entrenar los árboles de decisión, generando como resultado un ensamble de diferentes modelos. Sin embargo, *bagging* implica que los árboles son similares entre ellos dado que, ante la presencia de un predictor muy fuerte junto a otros predictores más moderados en un conjunto de datos, cada árbol que sea construído tomará el predictor fuerte en su nodo raíz.

Teniendo en cuenta esto, en el 2001 Leo Breiman de la Universidad de California, Berkeley, desarrolló una extensión de *bagging* que mejoró significativamente la precisión de clasificación [Breiman, 2001], conocido como *random forest* (Figura 6.2).

La metodología RF involucra dos pasos de aleatoridad en la formación de los árboles componentes:

- Primero, en la toma al azar de subconjuntos de datos de entrenamiento (selección con reemplazo). Así, para el entrenamiento de cada árbol se usa un subconjunto distinto de datos mientras que para la evaluación del modelo se usan los datos remanentes (conocidos como *out-of-bag*).

- Segundo, en cada partición el algoritmo considera solamente una fracción de las variables predictoras en vez de todas juntas. Esto resulta en árboles con diferentes predictores en la partición inicial y, por lo tanto, en árboles no correlacionados y un resultado promedio más confiable. Este incremento de diversidad hace que RF sea más robusto ante variables predictoras correlacionadas [Horning et al., 2010].

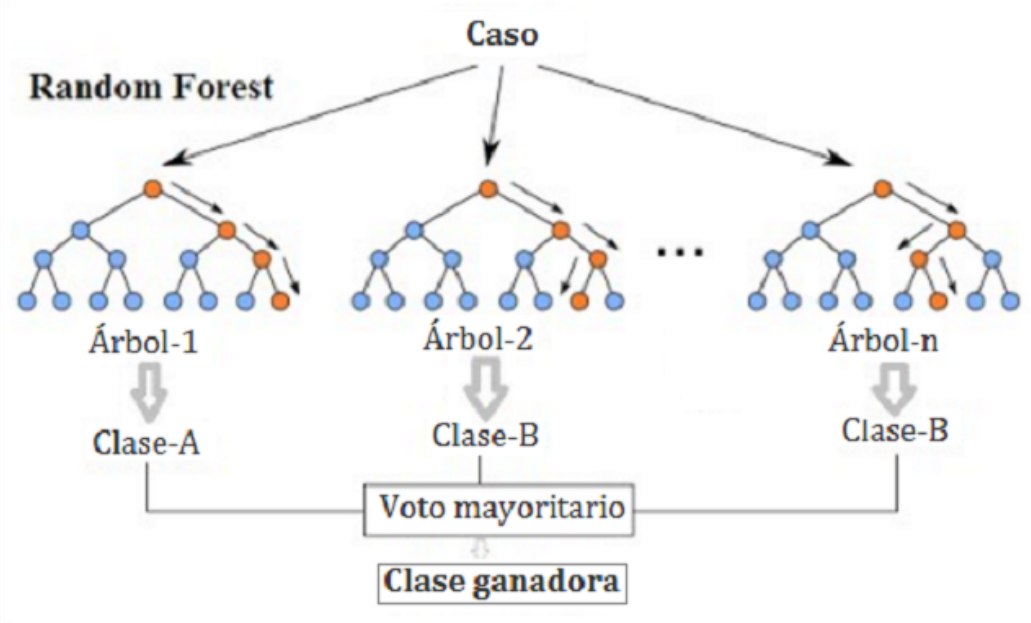

Figura 6.2: Esquema de *random forest*. El algoritmo construye una cantidad n de árboles de decisión en paralelo durante la etapa de entrenamiento. La salida o voto mayoritario del sistema es la moda de las clases (clasificación) o la predicción media (regresión).

Tanto en problemas de regresión como de clasificación, el desempeño de RF en una observación i generalmente es cuantificada mediante una función de pérdida, la cual mide la diferencia entre la respuesta verdadera y_i y la respuesta de predicción $\hat{y}_i$.

La métrica clásica y más directa es definida para una observación i como:

$$e_i = (y_i + \hat{y}_i)^2 = L(y_i, \hat{y}_i) \tag{6.1}$$

donde *L(.,.)* es la función de pérdida $L(x,y) = (x - y)^2$. En el caso de regresión, esto corresponde al error cuadrado.

El error cuadrado medio (MSE) es una métrica que mide el promedio de la ecuación 6.1, pero es sensible a *outliers*.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i + \hat{y}_i)^2 \tag{6.2}$$

La raíz cuadrada del error cuadrado medio (RMSE) es una métrica ampliamente usada en problemas de regresión, el rango va desde 0 hasta infinito y se apunta a valores bajos de esta métrica. Al igual que MSE, RMSE también es sensible a *outliers*.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i + \hat{y}_i)^2} \tag{6.3}$$

El error absoluto medio (MAE) es el promedio de los errores absolutos y representa una métrica robusta a los valores *outliers*. Conceptualmente, MAE es la métrica de evaluación más sencilla en problemas de regresión que indica qué tan alejadas estuvieron las predicciones.

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|x_i - x| \tag{6.4}$$

donde N es el número total de errores y $|x_i - x|$ son los errores absolutos.

En este capítulo se describen los resultados obtenidos luego de utilizar el algoritmo RF provisto por la plataforma *H2O.ai* sobre los subconjuntos de *Bacteroidetes* y *Firmicutes*. La idea principal es predecir una variable fundamental en estudios de la microbiota que caracteriza el grado de distribución de las bacterias en una muestra, como es la Diversidad. Para ello, se tendrá en cuenta las variables de metadatos y los valores de abundancia relativa de cada grupo de bacterias.

Una manera visual muy útil que ofrece *H2O.ai* para interpretar los resultados del modelo de predicción de diversidad es mediante la generación de una tabla con el grado de importancia de las variables. Esta representa la influencia relativa de cada

variable sobre la variable blanco o *target*: tiene en cuenta si una determinada variable fue seleccionada en cada paso de partición durante el proceso de construcción del árbol y cuánto mejoró (disminuyó) el error cuadrado (sobre todos los árboles).

Además, se decidió explorar cómo resultan las métricas de evaluación del modelo de regresión mediante el ajuste del número de árboles (N_{tree}). Se determinó de manera empírica el número de árboles para obtener un modelo de predicción con el menor error.

6.1. Bacteroidetes

Los resultados indicaron que la edad de los individuos fue el factor más importante en el modelo para predecir la diversidad de la microbiota intestinal. En cuanto a las abundancias de bacterias, se observó que dos especies tuvieron la mayor importancia relativa en la identificación de la diversidad: *Prevotella ruminicola* y *Bacteroides ovatus*. Por otra parte, de las restantes variables de metadata en el subconjunto de datos, el BMI de los individuos tuvo una posición destacable en el *ranking* de importancias, seguido por la Nacionalidad. Se observó que el género no fue importante en la predicción de la diversidad (Cuadro 6.1).

Mientras que la lista de valores de importancia para cada variable es informativa, se obtiene una mejor interpretación de cada una luego de escalar entre 0 y 1 y presentarlas en orden descendente de importancia (Figura 6.3).

En cuanto a las métricas de evaluación del modelo, se observó que comenzando con un valor inicial de $N_{tree} = 50$ y ascendiendo, MSE empezó a disminuír hasta mantenerse constante con $N_{tree} = 150$. A medida que aumentaba N_{tree}, los valores de RMSE disminuyeron hasta $N_{tree} = 250$. Los valores de MAE disminuyeron desde $N_{tree} = 50$ hasta $N_{tree} = 200$ (Cuadro 6.2).

Dado que MSE y RMSE son métricas sensibles a valores *outliers* y que estos existen en el subconjunto de datos (ver Figura 3.6), se decidió tener en cuenta los valores de MAE en el ajuste de parámetros, porque esta métrica es robusta. De esta manera, se determinó que el modelo de predicción de la diversidad biológica en *Bacteroidetes*

Variable	Importancia relativa	Importancia escalada
Edad	222,64	1,00
Prevotella ruminicola	187,33	0,84
BMI	153,52	0,68
Bacteroides ovatus	108,42	0,48
Nacionalidad	106,13	0,47
Bacteroides plebeius	83,35	0,37
Parabacteroides distasonis	77,08	0,34
Bacteroides stercoris	75,19	0,33
Bacteroides splachnicus	74,17	0,33
Allistipes	68,61	0,30
Prevotella tannerae	67,41	0,30
Tannerella	65,61	0,29
Prevotella oralis	64,48	0,28
Bacteroides vulgatus	64,10	0,28
Bacteroides uniformis	55,36	0,24
Bacteroides intestinalis	53,81	0,24
Bacteroides fragilis	49,97	0,22
Sexo	39,91	0,17
Prevotella melaninogenica	37,61	0,16

Cuadro 6.1: Listado de las variables predictoras sobre la diversidad e importancia relativa de las variables como resultado de la implementación de RF.

Número de árboles	MSE	RMSE	MAE	RMLE
50	0,040	0,201	0,153	0,030
100	0,039	0,199	0,152	0,029
150	0,038	0,196	0,151	0,029
200	0,038	0,196	0,150	0,029
250	0,038	0,195	0,150	0,029
300	0,038	0,196	0,151	0,029

Cuadro 6.2: Comparación de diversas métricas en la evaluación de RF de regresión dependiente del número de árboles en el subconjunto *Bacteroidetes*. MSE = *Mean Squared Error*, RMSE = *Root Mean Squared Error*.

mediante *random forest* es óptimo con $N_{tree} = 200$.

6.2. *Firmicutes*

Se observó que en el modelo de predicción de la diversidad de bacterias de *Firmicutes*, las variables de metadata fueron las más importantes, comenzando con el BMI de los individuos, seguida por la Nacionalidad y la Edad. En cuanto a las abundancias de los 74 tipos de bacterias presentes en este subconjunto, se observó que dos

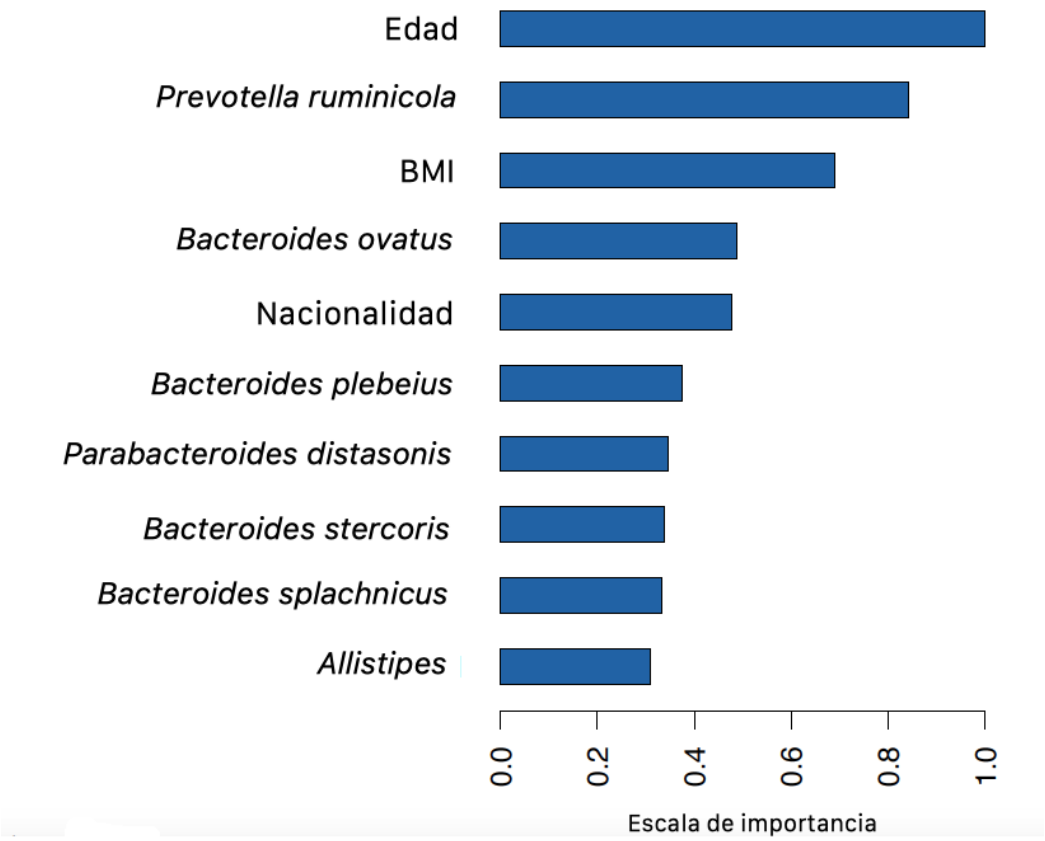

Figura 6.3: Escala de la importancia relativa de las diez primeras variables sobre la diversidad luego de implementar RF.

miembros tuvieron la mayor importancia relativa en la predicción de la diversidad: *Lachnobacillus bovis* y *Granulicatella* (Cuadro 6.3). Algunas de las restantes especies de *Firmicutes*, listadas en orden decreciente de importancia para el modelo, han sido asociadas a obesidad [Salonen et al., 2014, Rabot et al., 2016]. En la sección de Conclusiones de este capítulo se desarrollará más sobre estos resultados.

Cuadro 6.3: Importancia relativa de las variables predictoras sobre la Diversidad luego de la implementación de RF.

Variable	Importancia relativa	Importancia escalada
BMI	99,44	1,00
Nacionalidad	81,66	0,82
Edad	69,81	0,70
Lachnobacillus bovis	60,99	0,61
Granulicatella	54,89	0,55

(Sigue)

Variable	Importancia relativa	Importancia escalada
Anaerovorax odorimutans	53,56	0,53
Ruminococcus obeum	47,96	0,48
Sporobacter termitidis	45,32	0,45
Eubacterium biforme	33,27	0,33
Lactobacillus gasseri	33,02	0,33
Clostridium leptum	32,39	0,32
Bryantella formatexigens	31,99	0,32
Aneurinibacillus	31,81	0,31
Peptostreptococcus anaerobius	27,18	0,27
Anaerotruncus colihominis	26,16	0,26
Butyrivibrio crossotus	26,07	0,26
Oscillospira guillermondii	26,06	0,26
Phascolarctobacterium faecium	24,72	0,24
Clostridium stercorarium	23,72	0,23
Anaerofustis	22,61	0,22
Clostridium cellulosi	20,98	0,21
Eubacterium siraeum	20,88	0,21
RuminococcusLactaris	20,36	0,20
UCI I	19,80	0,19
Peptostreptococcus micros	19,59	0,19
Dorea formicigenerans	19,05	0,19
Eubacterium hallii	18,86	0,18
Lactobacillus catenaformis	17,69	0,17
Clostridium nexile	17,55	0,17
Anaerostipes caccae	17,41	0,17
Clostridium orbiscindens	16,15	0,16
Roseburia intestinalis	16,15	0,16

(Sigue)

Variable	Importancia relativa	Importancia escalada
Papillibacter cinnamivorans	16,10	0,16
Asteroleplasma	15,99	0,16
Peptococcus niger	15,96	0,16
Enterococcus	15,72	0,15
Bulleidia moorei	14,56	0,14
Lachnospira pectinoschiza	14,47	0,14
Megasphaera elsdenii	14,05	0,14
Eubacterium rectale	13,96	0,14
Ruminococcus gnavus	13,88	0,13
Megamonas hypermegale	13,76	0,13
Uncultured Selenomonadaceae	13,40	0,13
Clostridium symbiosum	12,49	0,12
Subdoligranulum variable	12,05	0,12
Clostridium difficile	12,02	0,12
UCI II	11,08	0,11
Faecalibacterium prausnitzii	10,74	0,10
Bacillus	10,68	0,10
Outgrouping Clostridium Cluster XIVa	10,66	0,10
Lactobacillus salivarius	10,63	0,10
Ruminococcus bromii	10,60	0,10
Coprococcus eutactus	10,33	0,10
Clostridium sensu stricto	10,27	0,10
Streptococcus intermedius	10,11	0,10
Sexo	10,10	0,10
Ruminococcus callidus	9,93	0,09
Clostridium ramosum	9,79	0,09
Streptococcus mitis	9,45	0,09

(Sigue)

Variable	Importancia relativa	Importancia escalada
EubacteriumCylindroides	9,33	0,09
Eubacterium limosum	9,23	0,09
Catenibacterium mitsuokai	9,18	0,09
Streptococcus bovis	9,11	0,09
Lactococcus	9,06	0,09
Aerococcus	8,66	0,08
Weissella	8,16	0,08
Coprobacillus catenaformis	7,92	0,07
Clostridium thermocellum	7,89	0,07
Staphylococcus	7,77	0,07
Dialister	7,53	0,07
Clostridium felsineum	7,49	0,07
Eubacterium ventriosum	7,18	0,07
Clostridium colinum	7,07	0,07
Clostridium sphenoides	7,05	0,07
Gemella	7,02	0,07
Lactobacillus plantarum	6,95	0,06
Veillonella	6,02	0,06
Mitsuokella multiacida	5,33	0,05

Se obtiene una mejor visualización de la influencia de las variables en el modelo de regresión luego de escalarlas entre 0 y 1 y presentarlas en orden decreciente de importancia (Figura 6.4).

En cuanto a las métricas de evaluación del modelo, se observó que con un valor de inicial de $N_{tree} = 50$ y ascendiendo, los valores de MSE, RMSE y MAE fueron variando levemente (Cuadro 6.4). Sin embargo, dado que MSE y RMSE no presentan grandes variaciones y que además son sensibles a *outliers*, se eligió a MAE como medida de

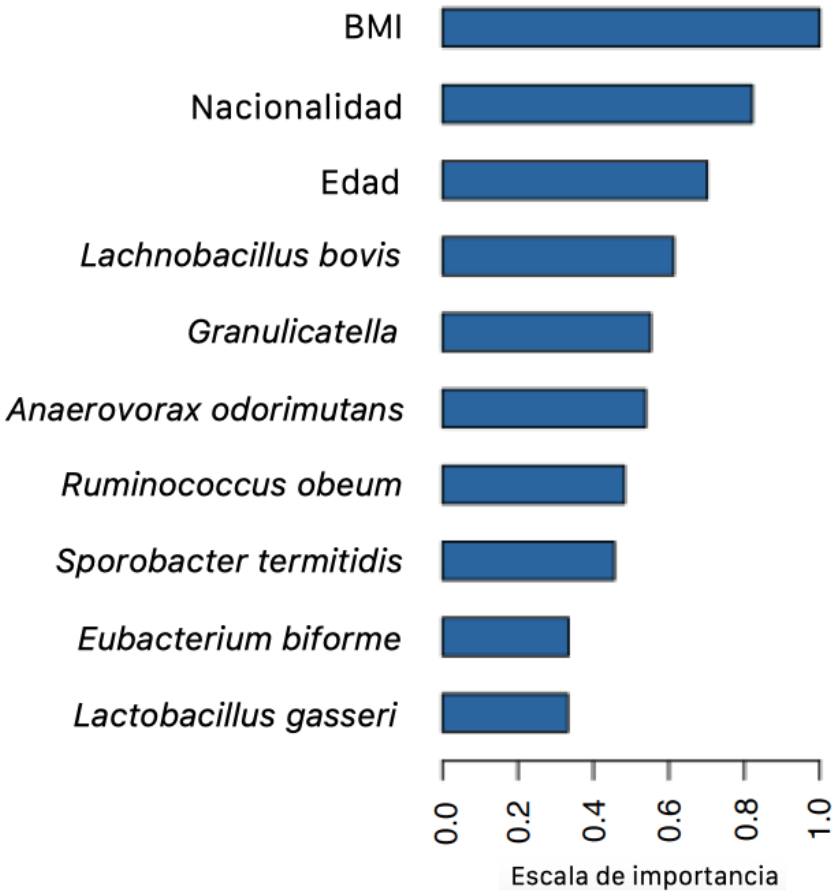

Figura 6.4: Escala de la importancia relativa de las diez primeras variables sobre la diversidad luego de implementar RF en el subconjunto de *Firmicutes*.

evaluación. Así, se determinó que el modelo de predicción de la diversidad biológica en *Firmicutes* mediante *random forest* es óptimo con $N_{tree} = 50$.

Número de árboles	MSE	RMSE	MAE	RMLE
50	0,030	0,175	0,136	0,026
100	0,031	0,176	0,136	0,026
150	0,031	0,176	0,137	0,026
200	0,030	0,175	0,137	0,026
250	0,030	0,175	0,137	0,026
300	0,030	0,176	0,137	0,026

Cuadro 6.4: Comparación de diversas métricas en la evaluación de RF de regresión dependiente del número de árboles en el subconjunto *Firmicutes*. MSE = *Mean Squared Error*, RMSE = *Root Mean Squared Error*.

6.3. Conclusiones

En esta parte de la tesis se utilizaron los dos subconjuntos de bacterias, *Bacteroidetes* y *Firmicutes*, para generar dos modelos de predicción de la diversidad en una población de individuos sanos mediante *random forest*. Independientemente de las métricas de evaluación en cada modelo de regresión, los resultados obtenidos sobre la importancia de determinadas variables son relevantes al momento de predecir la diversidad de la población.

Tanto la Edad y el BMI de los individuos aparecen como las dos variables más relevantes en los modelos generados para *Bacteroidetes* y *Firmicutes*. Estos resultados están avalados por trabajos previos que indican una importante influencia de estos factores en la dinámica del microbioma intestinal [Yatsunenko et al., 2012, Egert et al., 2006].

Con respecto a las bacterias, del subconjunto de 15 bacterias de *Bacteroidetes*, dos especies tuvieron mayor importancia en el modelo de predicción:

- *Prevotella ruminicola* pertenece a un género de bacterias cuya presencia en la microbiota intestinal determina las respuestas de individuos a suplementos dietarios [Chung et al., 2020].

- *Bacteroides ovatus* es una especie dominante en el intestino y ha sido identificada en trabajos previos como un probiótico de próxima generación debido a sus efectos preventivos en inflamación intestinal (revisado en [Tan et al., 2018]).

Del subconjunto de bacterias de *Firmicutes*, dos miembros tuvieron mayor importancia en el modelo de predicción: *Lachnobacillus bovis* y *Granulicatella*. Si bien no existen trabajos basados exclusivamente en estas dos especies, las abundancias relativas de las mismas están influenciadas por el tipo de dieta y uso de antibióticos en individuos obesos [Salonen et al., 2014, Reijnders et al., 2016].

Capítulo 7

Gradient Boosting Machine

Boosting es una técnica de ensamble que permite mejorar el rendimiento predictivo en problemas de clasificación y regresión, tal como árboles de decisión. Se basa en el agregado secuencial de nuevos modelos al ensamble que van mejorando levemente los errores remanentes. Los árboles de decisión se conectan secuencialmente (es decir, en serie) para obtener un modelo fuerte. Esta técnica usa una estrategia diferente de construcción de ensambles en comparación con la estrategia empleada en *random forest* descripta anteriormente.

Se comienza ajustando un modelo inicial a los datos. Luego, se construye un segundo modelo teniendo en cuenta los errores de predicción del primero y así sucesivamente se van agregando modelos débiles de manera secuencial, permitiendo que el algoritmo aprenda (Figura 7.1). Con cada modelo en el ensamble se van haciendo mejoras pequeñas e incrementales, lo cual evita el *overfitting* o sobreajuste de los datos.

De esta manera, *Gradient Boosting Machine* (GBM) construye un modelo aditivo de manera progresiva por etapas. En cada iteración se da un paso en la dirección de minimizar el error de predicción, en el espacio de posibles predicciones para cada caso de entrenamiento.

Además de la predicción de una variable dependiente en base a un conjunto de variables predictoras independientes es relevante también identificar las mismas y su grado de importancia en la predicción a fin de entender el proceso subyacente. Para ello, la implementación de GBM produce un *ranking* de las variables individuales

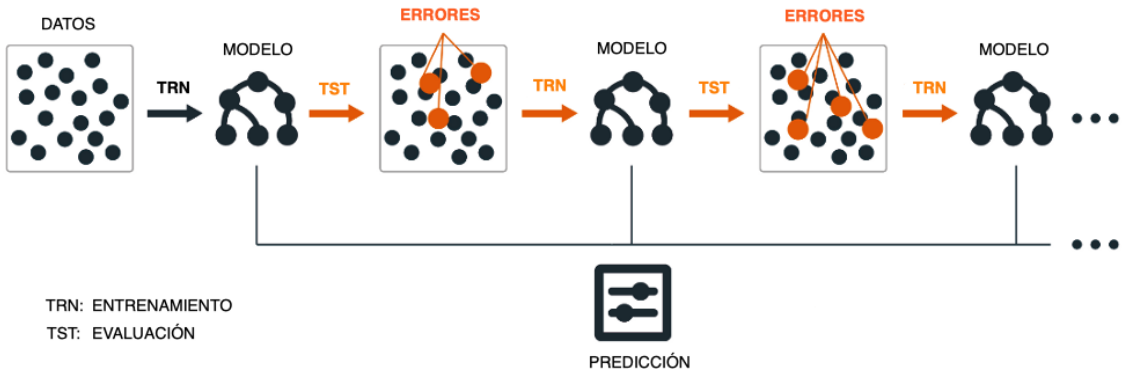

Figura 7.1: En GBM, cada modelo sucesivo intenta corregir los errores del conjunto combinado de todos los modelos anteriores. TRN=Entrenamiento, TST=Testeo.

representando la influencia relativa en el modelo.

En este capítulo se describen los resultados luego de usar el algoritmo *Gradient Boosting* provisto por la plataforma *H2O.ai* sobre los subconjuntos de *Bacteroidetes* y *Firmicutes*. El objetivo es predecir una variable fundamental en estudios de la microbiota que caracteriza el grado de distribución de las bacterias en una muestra, como es la Diversidad. Para ello, se tendrá en cuenta las variables de los metadatos y los valores de abundancia relativa de cada grupo de bacterias.

7.1. *Bacteroidetes*

Los resultados indicaron que el BMI y la edad de los individuos fueron los factores con mayor importancia en el modelo para predecir la diversidad de la microbiota intestinal. Con las restantes variables de metadatos, se observó que la Nacionalidad tuvo mayor relevancia que el Sexo. En cuanto a las especies de bacterias, se observó que *Bacteroides plebeius*, seguido por *Prevotella ruminicola* y *Bacteroides ovatus* fueron las que mayor importancia tuvieron en el modelo (Cuadro 7.1).

La representación gráfica de los diez primeros predictores que contribuyen en mayor medida a predecir la diversidad está indicada de manera descendiente en la Figura 7.2.

Variable	Importancia relativa	Importancia escalada
BMI	18,14	1,00
Edad	17,67	0,97
Nacionalidad	10,96	0,60
Bacteroides plebeius	10,92	0,60
Prevotella ruminicola	8,51	0,46
Bacteroides ovatus	7,00	0,38
Parabacteroides distasonis	6,26	0,34
Allistipes	5,01	0,27
Prevotella oralis	4,94	0,27
Bacteroides uniformis	4,84	0,26
Bacteroides fragilis	4,59	0,25
Bacteroides splachnicus	4,57	0,25
Prevotella tannerae	3,77	0,20
Sexo	2,81	0,15
Bacteroides vulgatus	2,42	0,13
Tannerella	2,06	0,11
Bacteroides stercoris	1,34	0,07
Prevotella melaninogenica	0,76	0,04
Bacteroides intestinalis	0,76	0,04

Cuadro 7.1: Listado de las variables predictoras sobre la Diversidad e importancia relativa de las variables como resultado de la implementación de GBM.

7.1.1. Ajuste de parámetros en el modelo

El ajuste de determinados parámetros apunta a la selección de un modelo con el menor error de predicción. Entre ellos, se decidió estudiar el impacto del número de árboles (N_{tree}) sobre diversas métricas que evalúan el error del modelo (MSE, RMSE, MAE, descriptas en las ecuaciones 6.2, 6.3 y 6.4).

Los resultados indicaron que a partir de $N_{tree} = 100$, la mejora en la disminución de los errores no fue evidente (Cuadro 7.2). De esta manera, el modelo de predicción de la diversidad biológica mediante *Gradient Boosting* es óptimo con $N_{tree} = 50$, porque se obtuvieron los valores más bajos de error.

7.2. *Firmicutes*

Luego de implementar el algoritmo de *Gradient Boosting*, utilizando las variables de metadatos y de bacterias, se obtuvo una lista de las variables más importantes para predecir la diversidad biológica desde el punto de vista del subconjunto de *Firmicutes*.

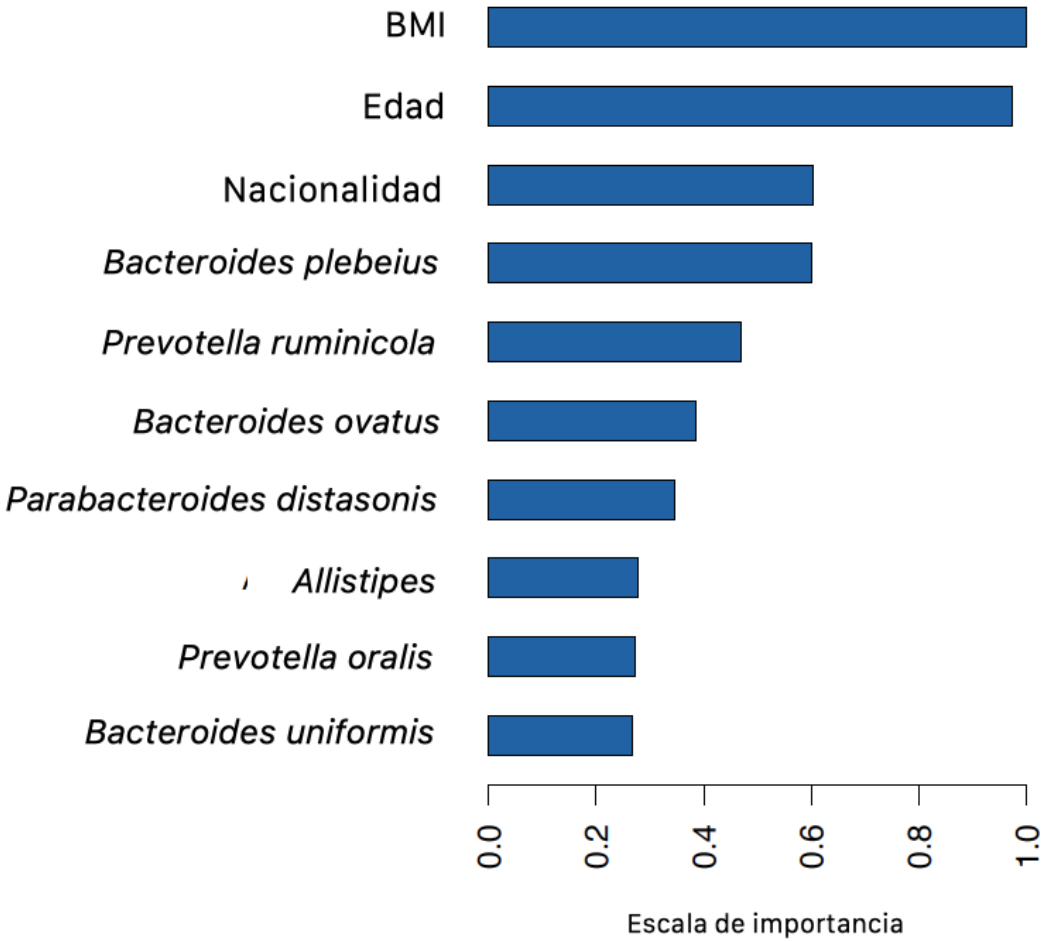

Figura 7.2: Escala de la importancia relativa de las diez primeras variables sobre la diversidad luego de implementar GBM usando el subconjunto de *Bacteroidetes*.

Número de árboles	MSE	RMSE	MAE	RMLE
50	0,038	0,195	0,147	0,029
100	0,039	0,197	0,147	0,029
150	0,039	0,199	0,150	0,029
200	0,039	0,198	0,150	0,029
250	0,039	0,197	0,148	0,029

Cuadro 7.2: Comparación de diversas métricas en la evaluación de GBM de regresión dependiente del número de árboles en el subconjuto *Bacteroidetes*. MSE = *Mean Squared Error*, RMSE = *Root Mean Squared Error*.

El BMI fue la variable con mayor importancia relativa sobre la diversidad, seguida por diversas especies como *Ruminococcus lactaris*, *Sporobacter termitidis*, *Bryantella formatexigens* y *Clostridium stercorarium*. De las restantes variables de metadatos, la Nacionalidad estuvo por encima de la edad. En cambio, la importancia de la variable

Sexo sobre la diversidad no tuvo una gran relevancia (Cuadro 7.3).

Cuadro 7.3: Importancia relativa de las variables predictoras sobre la Diversidad luego de la implementación de GBM.

Variable	Importancia relativa	Importancia escalada
BMI	11,74	1,00
Ruminococcus lactaris	11,04	0,94
Sporobacter termitidis	10,78	0,91
Bryantella formatexigens	10,45	0,89
Clostridium stercorarium	10,16	0,86
Nacionalidad	8,98	0,76
Lachnospira pectinoschiza	8,68	0,74
Eubacterium biforme	8,59	0,73
Clostridium leptum	8,34	0,71
Edad	7,92	0,67
UCI I	7,78	0,66
Eubacterium siraeum	6,99	0,59
Clostridium symbiosum	6,18	0,52
Faecalibacterium prausnitzii	4,88	0,41
Anaerotruncus colihominis	4,84	0,41
Peptococcus niger	3,43	0,29
Lachnobacillus bovis	3,35	0,28
Ruminococcus obeum	2,64	0,22
Anaerovorax odorimutans	2,55	0,21
Eubacterium rectale	2,19	0,18
Clostridium nexile	2,08	0,17
Granulicatella	1,86	0,15
Lactobacillus gasseri	1,74	0,14
Oscillospira guillermondii	1,70	0,14
Aneurinibacillus	1,34	0,11

(Sigue)

Variable	Importancia relativa	Importancia escalada
Eubacterium hallii	1,32	0,11
Megasphaera elsdenii	1,31	0,11
Roseburia intestinalis	1,10	0,09
UCI II	1,00	0,08
Butyrivibrio crossotus	0,96	0,08
Phascolarctobacterium faecium	0,95	0,08
Papillibacter cinnamivorans	0,92	0,07
Lactococcus	0,92	0,07
Dorea formicigenerans	0,82	0,07
Clostridium orbiscindens	0,78	0,06
Lactobacillus plantarum	0,72	0,06
Anaerofustis	0,70	0,06
Clostridium cellulosi	0,69	0,05
Lactobacillus catenaformis	0,63	0,05
Streptococcus bovis	0,51	0,04
Lactobacillus salivarius	0,51	0,04
Sexo	0,48	0,04
Ruminococcus gnavus	0,42	0,03
Clostridium sphenoides	0,39	0,03
Bulleidia moorei	0,38	0,03
Clostridium sensu stricto	0,33	0,02
Eubacterium ventriosum	0,29	0,02
Megamonas hypermegale	0,27	0,02
Bacillus	0,26	0,02
Streptococcus intermedius	0,25	0,02
Coprococcus eutactus	0,22	0,01
Aerococcus	0,20	0,01

(Sigue)

Variable	Importancia relativa	Importancia escalada
Mitsuokella multiacida	0,19	0,01
Staphylococcus	0,18	0,01
Ruminococcus bromii	0,17	0,01
Enterococcus	0,17	0,01
Anaerostipes caccae	0,16	0,01
Clostridium difficile	0,15	0,01
Clostridium colinum	0,13	0,01
Outgrouping Clostridium ClusterXIVa	0,10	–
Gemella	0,07	–
Ruminococcus callidus	0,04	–
Coprobacillus catenaformis	0,03	–
Asteroleplasma	–	–
Catenibacterium mitsuokai	–	–
Clostridium felsineum	–	–
Clostridium ramosum	–	–
Clostridium thermocellum	–	–
Dialister	–	–
Eubacterium cylindroides	–	–
Eubacterium limosum	–	–
Peptostreptococcus anaerobius	–	–
Peptostreptococcus micros	–	–
Streptococcus mitis	–	–
Subdoligranulum variable	–	–
Uncultured Selenomonadaceae	–	–
Veillonella	–	–
Weissella	–	–

La representación gráfica de los diez primeros predictores que contribuyen en mayor importancia a la diversidad se muestran de manera descendiente en la Figura 7.3.

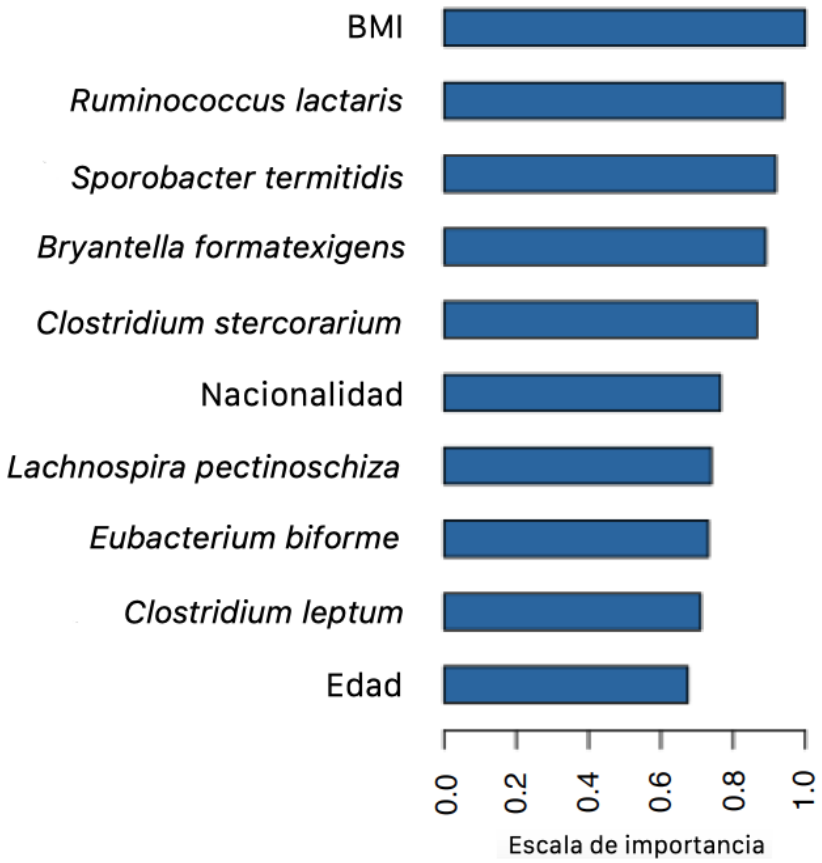

Figura 7.3: Escala de la importancia relativa de las diez primeras variables sobre la diversidad luego de implementar GBM en el subconjunto de *Firmicutes*.

7.2.1. Ajuste de parámetros en el modelo

Se evaluó el impacto del número de árboles (N_{tree}) sobre diversas métricas que evalúan el error del modelo (MSE, RMSE, MAE, descriptas en las ecuaciones 6.2, 6.3 y 6.4), para así apuntar a la selección de un modelo con el menor error de predicción.

Se observó que a partir de $N_{tree} = 200$ se llegó a una disminución de los errores (Cuadro 7.4). De esta manera, el modelo de predicción de la diversidad biológica en *Firmicutes* mediante GBM es óptimo con $N_{tree} = 200$, que corresponde a los valores más bajos de error.

Número de árboles	MSE	RMSE	MAE	RMLE
50	0,018	0,135	0,107	0,020
100	0,014	0,121	0,096	0,017
150	0,014	0,118	0,093	0,017
200	0,013	0,117	0,093	0,017
250	0,013	0,117	0,093	0,017

Cuadro 7.4: Comparación de diversas métricas en la evaluación de GBM de regresión dependiente del número de árboles en el subconjuto *Firmicutes*. MSE = *Mean Squared Error*, RMSE = *Root Mean Squared Error*.

7.3. Conclusiones

En esta capítulo se utilizaron los dos subconjuntos de bacterias, *Bacteroidetes* y *Firmicutes*, para generar dos modelos de predicción de la diversidad en una población de individuos sanos mediante *gradient boosting*.

Ambos modelos coincidieron en la variable BMI como la de mayor importancia relativa para predecir la diversidad.

Sin embargo, el modelo de predicción de *Bacteroidetes* identificó a la edad y a la nacionalidad como las siguientes variables en importancia. Con respecto a las bacterias, dos especies tuvieron mayor importancia en el modelo:

- *Bacteroides plebeius* se encuentra presente en el intestino de ciertas poblaciones de individuos y confieren una ventaja metabólica al hospedador. Esto se debe a que es un ejemplo de cómo la microbiota residente puede adquirir genes nuevos funcionales como una manera de adaptación mediante un proceso conocido como transferencia lateral de genes. Así *B. plebeius* adquirió el gen de la enzima porfirinasa a partir de una bacteria asociada a algas rojas marinas, las cuales se usan en la preparación de sushi. El beneficio para el hospedador reside en que la microbiota intestinal puede así extraer energía de los polisacáridos complejos presentes [Hehemann et al., 2010].

- *Prevotella ruminicola*, sobre la cual ya se comentó su relevancia en la microbiota intestinal en el Capítulo 7.

En cuanto al modelo generado a partir del subconjunto *Firmicutes*, las siguientes

variables con importancia sobre la diversidad corresponden a las bacterias tal como *Ruminococcus lactaris* y *Sporobacter termitidis*.

- Es interesante destacar que las especies del género *Ruminococcus* forman un grupo robusto en la microbiota intestinal humana: en Arumugam *et al* se analizaron 33 muestras de diferentes poblaciones y, en base a géneros discriminativos, se identificó el predominio de tres *clusters*, a los que se denominó enterotipos. El enterotipo 3 está enriquecido en *Ruminococcus*, cuyas especies están principalmente especializadas en degradar y convertir azúcares complejos en una variedad de nutrientes para el hospedador [Arumugam et al., 2011, La Reau and Suen, 2018].

- *Sporobacter termitidis* es una bacteria asociada a la mucosa intestinal y con probable efecto antiinflamatorio, tal como indica el reporte de un caso de transplante de microbiota fecal (FMT, por sus siglas en inglés) en un individuo con infección recurrente gastrointestinal. Luego del tratamiento de FMT, se observó un aumento sostenido en la abundancia intestinal de algunas bacterias, entre las cuales se destacaba *Sporobacter termitidis* [Satokari et al., 2014]. La relevancia de esta bacteria también se destacó en la microbiota intestinal de personas sanas en comparación con pacientes de enfermedad intestinal inflamatoria.

Capítulo 8

Discusión

A lo largo del tracto gastrointestinal se aloja un vasto sistema de millones de bacterias, hongos y virus, cuyas comunidades residentes conviven en un balance complejo entre sí y con el hospedador. La composición de las miles de especies de bacterias en el microbioma intestinal humano es muy variable. Tal diversidad depende de factores como la dieta, el ambiente, la genética del hospedador, la edad, entre otros y su estudio representa entonces un desafío importante para las ciencias biomédicas debido a que desbalances en la microbiota intestinal están asociados a enfermedades.

Los avances tecnológicos hicieron posible que la información biológica pueda ser cuantificada en paralelo y a escala masiva, impulsando a que las áreas de las biociencias se enfrenten a cantidades cada vez mayores de datos ómicos (referidos al conjunto total de biomoléculas como ADN, ARN, proteínas, metabolitos, etc). Dentro de la genómica, una de las primeras tecnologías que ha permitido el análisis masivo y en paralelo de la expresión de genes es el *microarray*. En esta tesis se usaron datos generados en *microarrays* filogenéticos que permiten evaluar la presencia y abundancia relativa de miles de bacterias de la microbiota intestinal proveniente de individuos sanos mediante la cuantificación de un gen marcador, como es el 16S rRNA.

Los microorganismos que componen la microbiota intestinal humana están clasificados taxonómicamente en seis grandes grupos o *phyla*, que contienen a su vez cientos de miembros. En esta tesis se abordó el análisis de dos de estos grupos, *Bacteroidetes* y *Firmicutes*, mediante el uso de algoritmos de aprendizaje automático. Si bien el

conjunto de datos utilizado ya ha sido utilizado previamente [Lahti et al., 2014], el enfoque del presente trabajo fue caracterizar la abundancia relativa de bacterias de los dos grupos mencionados con nuevas herramientas de análisis para evaluar posibles asociaciones con información adicional de los individuos.

El conjunto de datos cuenta con información sobre el índice de masa corporal (BMI) de los individuos, además del sexo, nacionalidad y edad. La decisión inicial de enfocar el estudio en *Bacteroidetes* y *Firmicutes* se basó en el hecho de que varios géneros de ambos grupos son predominantes en el microbioma intestinal. Además existe una relación con la obesidad, tanto en humanos como en modelos animales, la microbiota de individuos obesos está mayormente enriquecida con bacterias del grupo *Firmicutes* y en menor medida las del grupo *Bacteroidetes* (Ley et al, 2006; Castaner et al, 2018).

Se buscó en principio determinar si es posible identificar bacterias con predominancia de tal manera que representen una gran proporción de la variabilidad. Es decir, reducir el número de variables a unas pocas para empezar a tener una idea sobre la dinámica de los datos. Para ello, PCA a menudo es usado como una herramienta de análisis exploratorio para reducir la dimensión de variables antes de armar modelos de predicción. Puede ser usado para reducir una gran cantidad de variables predictoras en un número menor de componentes principales, en particular si se aplica en conjuntos de datos no muy limpios o con variables fuertemente correlacionadas. En ambos subconjuntos de datos no fue posible explicar un porcentaje de variabilidad importante sin considerar un gran número de componentes. Aunque menor a la cantidad original en cada caso, el no poder reducir a dos a tres componentes (como suele ocurrir en conjuntos de datos ideales) implica que la variabilidad no está dominada por unas pocas bacterias sino que justamente revela el comportamiento complejo y dinámico de la comunidad ecológica.

Un aspecto de gran relevancia en la minería de datos ómicos al momento de presentar resultados es el análisis visual. Por ello, se eligió implementar el algoritmo de SOM (mapas auto-organizados) que usa como modelo el espacio de vectores para re-

presentar datos en mapas de dos dimensiones: Cada valor a través de N muestras podría ser referido como un punto de dato en un espacio de N dimensiones. Miles de puntos de datos por lo tanto formarían nubes de datos en el espacio, con una estructura topológica intrínseca debido a las relaciones geométricas. De esto se desprende que a mayor similitud en el nivel de valor del dato, más cercano es el espacio geométrico que ocupan. Tal preservación topológica es de particular importancia en la fase exploratoria de la minería de datos ómicos dado que generalmente no se tiene un conocimiento *a priori* de la estructura de los mismos.

Los resultados mostraron que cada capa o mapa representa un tipo de dato: un mapa bidimensional para cada variable de metadatos (BMI, nacionalidad, edad, sexo y diversidad) del mismo modo que un mapa para cada bacteria (cuya abundancia relativa está representada en niveles de expresión del gen 16S rRNA). La interpretación de los resultados de SOM es mediante la exploración de las relaciones geométricas entre los nodos (dado que cada mapa preserva la forma y densidad) lo que favorece la identificación más directa de similitudes y diferencias entre las capas.

Por ejemplo, mediante este tipo de exploración se encontró una relación entre nodos que agrupaban individuos (principalmente mujeres) con sobrepeso/obesos y abundancias altas de solamente dos bacterias, *Peptostreptococcus micros* y *Tannerella*. Es interesante destacar que ambas especies, cuando están presentes en niveles altos, han sido asociadas previamente a roles patógenos en enfermedades periodontales (procesos infecciosos que afectan a las estructuras de soporte de los dientes) [Nagarajan et al., 2018]. Teniendo en cuenta que existen estudios que investigan el comportamiento del microbioma oral durante una infección periodontal y su influencia en complicaciones de salud, como diabetes, enfermedad cardiovascular y obesidad [Goodson et al., 2009], la importancia de los resultados obtenidos brinda una base para futuros estudios sobre el posible rol de bacterias como marcadores biológicos de una condición de sobrepeso en desarrollo.

Sin dejar de tener en cuenta que la relación entre la obesidad y el microbioma es compleja y variable, el tipo de visualización de resultados que ofrecen los mapas auto-

organizados representa un elemento que podría sumar en el análisis de una población en estudio dentro del marco de un proyecto biomédico.

El análisis de la diversidad de microorganismos que habitan en el cuerpo humano es fundamental para comprender la estructura, biología y ecología de las comunidades de los mismos; dicho análisis representa un primer componente crítico en el estudio de la microbiota.

Aún cuando en condiciones de salud se mantiene un núcleo central de bacterias relativamente estable en el tracto gastrointestinal a lo largo de la etapa adulta, la diversidad bacteriana depende de un número de factores tanto intrínsecos como extrínsecos del hospedador.

Por lo tanto, para determinar si era posible predecir la diversidad biológica del sistema se usaron dos algoritmos de aprendizaje supervisado: *Random Forest* (RF)y *Gradient Boosting Modeling* (GBM). Ambos son métodos de ensamble, basados en entrenar múltiples modelos débiles (*weak learners*) para resolver el mismo problema y así generar un modelo fuerte con mejor desempeño que sus componentes individuales. Sin embargo, difieren en dos puntos importantes: la forma en que se definen los datos a ser entrenados y el orden de entrenamiento de los modelos débiles.

La decisión de usar un algoritmo u otro depende claramente de las características del conjunto de datos. En líneas generales, existen diferencias en cuanto a que las técnicas de *bagging* (*random forest*) apuntan a disminuir la variabilidad típica de los modelos débiles, tal como los árboles de decisión. Mientras que las técnicas de *boosting* apuntan a reducir el sesgo de los modelos débiles mediante la disminución del error de predicción en cada paso.

Luego de entrenar los datos con RF o GBM es posible tener una idea de las variables con mayor influencia en el modelo. Para una variable numérica continua, como es la diversidad, la importancia relativa de cada variable es calculada en base a la reducción de la suma de los errores cuadrados cada vez que una determinada variable es elegida para partición.

En los modelos generados tanto por RF como GBM para el subconjunto de *Bac-*

teroidetes se observó que las primeras cinco variables más importantes incluyeron a la edad, el BMI y la nacionalidad. Y en cuanto a la abundancia de bacterias, *Prevotella ruminicola* es la que se destacó en ambos modelos. Es decir, el uso de ambos métodos de ensamble mostró resultados similares en cuanto a la influencia relativa de un conjunto de variables en la diversidad del microbioma intestinal.

La progresión de la edad del hospedador es un factor importante sobre la diversidad de la microbiota intestinal. Con la edad, las funciones beneficiosas que aporta una microbiota intestinal saludable empiezan a disminuir y empieza a aumentar, en cambio, la frecuencia de procesos inflamatorios y enfermedad especialmente en las personas de edad avanzada [Nagpal et al., 2018, Xu et al., 2019].

En cuanto a la influencia del BMI, diversos estudios que comparan la microbiota intestinal entre individuos obesos y delgados indican que la variación en el grado de diversidad está asociada al peso corporal: los individuos obesos (BMI alto) presentan una diversidad baja [Turnbaugh et al., 2009, Bäckhed et al., 2012].

El origen geográfico de un individuo (representado en el conjunto de datos a través de la variable nacionalidad) involucra una cantidad de factores ambientales que ejercen un efecto específico en la adquisición, composición y estabilidad de la microbiota intestinal. Varios trabajos muestran las diferencias en composición y diversidad de las microbiotas intestinales humanas en individuos saludables a nivel poblacional [Nishijima et al., 2016, Gupta et al., 2017].

El hecho de que dos métodos supervisados distintos hayan mostrado resultados similares para un mismo subconjunto de datos destaca la influencia ya conocida de ciertas variables sobre la diversidad de la microbiota intestinal.

Capítulo 9

Conclusiones

La evaluación de la composición de la microbiota intestinal humana facilita en gran medida la comprensión de su papel en la fisipatología humana. El conocimiento de los diversos factores que influyen sobre la composición, dinámica y funciones de la microbiota podría servir como una herramienta complementaria en la práctica médica.

La creciente investigación del microbioma intestinal y su influencia en la salud y enfermedad ha acelerado la necesidad de una integración de campos multidisciplinarios en su análisis. En general, los profesionales de la salud no están preparados para trabajar en todos los pasos del proceso de análisis de datos (limpieza, filtrado, elección de algoritmos, interpretación, etc.). Por lo tanto, la ciencia de datos y el aprendizaje automático pueden contribuir a la traducción de resultados innovadores en conocimientos valiosos que brinden apoyo a la toma de decisiones, algo muy requerido en la **medicina de precisión**.

Una contribución de esta tesis es brindar un estudio del microbioma intestinal humano desde el punto de vista del aprendizaje automático. Es decir, no se enfocó en armar flujos de trabajo bioinformáticos ampliamente usados en la investigación en esta área, sino que intenta brindar herramientas que pueden complementar al análisis bioinformático.

El análisis usando mapas auto-organizados brinda una visión global más intuitiva e informativa del comportamiento de unos pocos módulos bien definidos de ítems correlacionados en comparación con un descubrimiento aislado de los niveles de expresión

de cientos o miles de ítems individuales.

Es probable que la abundancia relativa de ciertas especies de *Prevotella* y *Bacteroides* representen un factor a tener en cuenta como biomarcadores en el ecosistema de al microbiota intestinal.

Bibliografía

[h2o, 2020] (2020). *h2o: R Interface for H2O*. R package version 3.30.0.6.

[Alberts et al., 2002] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., et al. (2002). Molecular biology of the cell, 6th edn new york. *NY: Garland Science*.

[Arumugam et al., 2011] Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., et al. (2011). Enterotypes of the human gut microbiome. *nature*, 473(7346):174–180.

[Bäckhed et al., 2004] Bäckhed, F., Ding, H., Wang, T., Hooper, L. V., Koh, G. Y., Nagy, A., Semenkovich, C. F., and Gordon, J. I. (2004). The gut microbiota as an environmental factor that regulates fat storage. *Proceedings of the national academy of sciences*, 101(44):15718–15723.

[Bäckhed et al., 2012] Bäckhed, F., Fraser, C. M., Ringel, Y., Sanders, M. E., Sartor, R. B., Sherman, P. M., Versalovic, J., Young, V., and Finlay, B. B. (2012). Defining a healthy human gut microbiome: current concepts, future directions, and clinical applications. *Cell host & microbe*, 12(5):611–622.

[Bäckhed et al., 2005] Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A., and Gordon, J. I. (2005). Host-bacterial mutualism in the human intestine. *science*, 307(5717):1915–1920.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[Burgos et al., 2021] Burgos, V., Piñero, T., Fernández, M. L., and Risk, M. (2021). A complementary approach in the analysis of the human gut microbiome applying self-organizing maps and random forest. In *International Conference on Applied Informatics*, pages 97–110. Springer.

[Chaban et al., 2006] Chaban, B., Ng, S. Y., and Jarrell, K. F. (2006). Archaeal habitats—from the extreme to the ordinary. *Canadian journal of microbiology*, 52(2):73–116.

[Chung et al., 2020] Chung, W. S. F., Walker, A. W., Bosscher, D., Garcia-Campayo, V., Wagner, J., Parkhill, J., Duncan, S. H., and Flint, H. J. (2020). Relative abundance of the prevotella genus within the human gut microbiota of elderly volunteers determines the inter-individual responses to dietary supplementation with wheat bran arabinoxylan-oligosaccharides. *BMC microbiology*, 20(1):1–14.

[Crovesy et al., 2020] Crovesy, L., Masterson, D., and Rosado, E. L. (2020). Profile of the gut microbiota of adults with obesity: A systematic review. *European Journal of Clinical Nutrition*, pages 1–12.

[De Filippo et al., 2010] De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S., Collini, S., Pieraccini, G., and Lionetti, P. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from europe and rural africa. *Proceedings of the National Academy of Sciences*, 107(33):14691–14696.

[Del Chierico et al., 2015] Del Chierico, F., Ancora, M., Marcacci, M., Camma, C., Putignani, L., and Conti, S. (2015). Choice of next-generation sequencing pipelines. In *Bacterial Pangenomics*, pages 31–47. Springer.

[Dominguez-Bello et al., 2010] Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., and Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habi-

tats in newborns. *Proceedings of the National Academy of Sciences*, 107(26):11971–11975.

[Donaldson et al., 2016] Donaldson, G. P., Lee, S. M., and Mazmanian, S. K. (2016). Gut biogeography of the bacterial microbiota. *Nature Reviews Microbiology*, 14(1):20–32.

[Dryad Digital Repository, 2021] Dryad Digital Repository (2021). Dryad digital repository. `https://datadryad.org/stash`, Last accessed on 2021-11-08.

[Dutta et al., 2019] Dutta, S. K., Verma, S., Jain, V., Surapaneni, B. K., Vinayek, R., Phillips, L., and Nair, P. P. (2019). Parkinson's disease: the emerging role of gut dysbiosis, antibiotics, probiotics, and fecal microbiota transplantation. *Journal of neurogastroenterology and motility*, 25(3):363.

[Egert et al., 2006] Egert, M., de Graaf, A. A., Smidt, H., de Vos, W. M., and Venema, K. (2006). Beyond Diversity: Functional Microbiomics of the Human Colon. *Trends Microbiol*, 14(2):86–91.

[Eloe-Fadrosh and Rasko, 2013] Eloe-Fadrosh, E. A. and Rasko, D. A. (2013). The human microbiome: from symbiosis to pathogenesis. *Annual review of medicine*, 64:145–163.

[European Society Neurogastroenterology and Motility, 2020] European Society Neurogastroenterology and Motility (2020). Comensal (bacteria) - editado por european society of neurogastroenterology and motility. `https://www.gutmicrobiotaforhealth.com/es/glossary/comensal-bacteria/`, Last accessed on 2021-11-08.

[Floch et al., 2016] Floch, M. H., Ringel, Y., and Walker, W. A. (2016). *The microbiota in gastrointestinal pathophysiology: implications for human health, prebiotics, probiotics, and dysbiosis*. Academic Press.

[Gandía-Aguiló et al., 2017] Gandía-Aguiló, V., Cibrián, R., Soria, E., Serrano, A.-J., Aguiló, L., Paredes, V., and Gandía, J.-L. (2017). Use of self-organizing maps

for analyzing the behavior of canines displaced towards midline under interceptive treatment. *Medicina oral, patologia oral y cirugia bucal*, 22(2):e233.

[Gill et al., 2006] Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *science*, 312(5778):1355–1359.

[Goodson et al., 2009] Goodson, J., Groppo, D., Halem, S., and Carpino, E. (2009). Is obesity an oral bacterial disease? *Journal of dental research*, 88(6):519–523.

[Greenhalgh et al., 2016] Greenhalgh, K., Meyer, K. M., Aagaard, K. M., and Wilmes, P. (2016). The human gut microbiome in health: establishment and resilience of microbiota over a lifetime. *Environmental microbiology*, 18(7):2103–2116.

[Gupta et al., 2017] Gupta, V. K., Paul, S., and Dutta, C. (2017). Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. *Frontiers in microbiology*, 8:1162.

[Han et al., 2019] Han, L., Yang, G., Dai, H., Yang, H., Xu, B., Li, H., Long, H., Li, Z., Yang, X., and Zhao, C. (2019). Combining self-organizing maps and biplot analysis to preselect maize phenotypic components based on uav high-throughput phenotyping platform. *Plant methods*, 15(1):1–16.

[Hehemann et al., 2010] Hehemann, J.-H., Correc, G., Barbeyron, T., Helbert, W., Czjzek, M., and Michel, G. (2010). Transfer of carbohydrate-active enzymes from marine bacteria to japanese gut microbiota. *Nature*, 464(7290):908–912.

[Horgan and Kenny, 2011] Horgan, R. P. and Kenny, L. C. (2011). 'omic'technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist*, 13(3):189–195.

[Horning et al., 2010] Horning, N. et al. (2010). Random forests: An algorithm for image classification and generation of continuous fields data sets. In *Proceedings of*

the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences, Osaka, Japan, volume 911.

[Hug et al., 2016] Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., et al. (2016). A new view of the tree of life. *Nature microbiology*, 1(5):16048.

[Koenig et al., 2011] Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., Angenent, L. T., and Ley, R. E. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4578–4585.

[Koliada et al., 2017] Koliada, A., Syzenko, G., Moseiko, V., Budovska, L., Puchkov, K., Perederiy, V., Gavalko, Y., Dorofeyev, A., Romanenko, M., Tkach, S., et al. (2017). Association between body mass index and firmicutes/bacteroidetes ratio in an adult ukrainian population. *BMC microbiology*, 17(1):120.

[Kowarik and Templ, 2016] Kowarik, A. and Templ, M. (2016). Imputation with the R package VIM. *Journal of Statistical Software*, 74(7):1–16.

[La Reau and Suen, 2018] La Reau, A. J. and Suen, G. (2018). The ruminococci: key symbionts of the gut ecosystem. *journal of microbiology*, 56(3):199–208.

[Lahti et al., 2014] Lahti, L., Salojärvi, J., Salonen, A., Scheffer, M., and De Vos, W. M. (2014). Tipping elements in the human intestinal ecosystem. *Nature communications*, 5(1):1–10.

[Lai et al., 2018] Lai, K., Twine, N., O'brien, A., Guo, Y., and Bauer, D. (2018). Artificial intelligence and machine learning in bioinformatics. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 55:272.

[Larsbrink et al., 2014] Larsbrink, J., Rogers, T. E., Hemsworth, G. R., McKee, L. S., Tauzin, A. S., Spadiut, O., Klinter, S., Pudlo, N. A., Urs, K., Koropatkin, N. M., et al. (2014). A discrete genetic locus confers xyloglucan metabolism in select human gut bacteroidetes. *Nature*, 506(7489):498–502.

[Leskovec et al., 2014] Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). *Bioinformatics Basics. Applications in Biological Science and Medicine*.

[Liu et al., 2019] Liu, H., Han, M., Li, S. C., Tan, G., Sun, S., Hu, Z., Yang, P., Wang, R., Liu, Y., Chen, F., et al. (2019). Resilience of human gut microbial communities for the long stay with multiple dietary shifts. *Gut*, 68(12):2254–2255.

[Lloyd-Price et al., 2016] Lloyd-Price, J., Abu-Ali, G., and Huttenhower, C. (2016). The healthy human microbiome. *Genome medicine*, 8(1):1–11.

[Matsuoka and Kanai, 2015] Matsuoka, K. and Kanai, T. (2015). The gut microbiota and inflammatory bowel disease. In *Seminars in immunopathology*, volume 37, pages 47–55. Springer.

[Morgan et al., 2013] Morgan, X. C., Segata, N., and Huttenhower, C. (2013). Biodiversity and functional genomics in the human microbiome. *Trends in genetics*, 29(1):51–58.

[Nagarajan et al., 2018] Nagarajan, M., Prabhu, V. R., and Kamalakkannan, R. (2018). Metagenomics: implications in oral health and disease. In *Metagenomics*, pages 179–195. Elsevier.

[Nagpal et al., 2018] Nagpal, R., Mainali, R., Ahmadi, S., Wang, S., Singh, R., Kavanagh, K., Kitzman, D. W., Kushugulova, A., Marotta, F., and Yadav, H. (2018). Gut microbiome and aging: Physiological and mechanistic insights. *Nutrition and healthy aging*, 4(4):267–285.

[National Human Genome Research Institute, 2020] National Human Genome Research Institute (2020). Bioinformática. `https://www.genome.gov/es/genetics-glossary/Bioinformatica`, Last accessed on 2021-11-08.

[Nishijima et al., 2016] Nishijima, S., Suda, W., Oshima, K., Kim, S.-W., Hirose, Y., Morita, H., and Hattori, M. (2016). The gut microbiome of healthy japanese and its microbial and functional uniqueness. *DNA Research*, 23(2):125–133.

[Rabot et al., 2016] Rabot, S., Membrez, M., Blancher, F., Berger, B., Moine, D., Krause, L., Bibiloni, R., Bruneau, A., Gérard, P., Siddharth, J., et al. (2016). High fat diet drives obesity regardless the composition of gut microbiota in mice. *Scientific reports*, 6(1):1–11.

[Ravel et al., 2011] Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S., McCulle, S. L., Karlebach, S., Gorle, R., Russell, J., Tacket, C. O., et al. (2011). Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4680–4687.

[Reese and Dunn, 2018] Reese, A. and Dunn, R. (2018). Drivers of Microbiome Biodiversity: A Review of General Rules, Feces, and Ignorance. *mBio*, 9(4):e01294–18.

[Reijnders et al., 2016] Reijnders, D., Goossens, G. H., Hermes, G. D., Neis, E. P., van der Beek, C. M., Most, J., Holst, J. J., Lenaerts, K., Kootte, R. S., Nieuwdorp, M., et al. (2016). Effects of gut microbiota manipulation by antibiotics on host metabolism in obese humans: a randomized double-blind placebo-controlled trial. *Cell metabolism*, 24(1):63–74.

[Ruan et al., 2011] Ruan, X., Gao, Y., Song, H., Chen, J., et al. (2011). A new dynamic self-organizing method for mobile robot environment mapping. *Journal of Intelligent Learning Systems and Applications*, 3(04):249.

[Ruggles et al., 2018] Ruggles, K. V., Wang, J., Volkova, A., Contreras, M., Noya-Alarcon, O., Lander, O., Caballero, H., and Dominguez-Bello, M. G. (2018). Changes in the gut microbiota of urban subjects during an immersion in the traditional diet and lifestyle of a rainforest village. *mSphere*, 3(4):e00193–18.

[Salonen et al., 2014] Salonen, A., Lahti, L., Salojärvi, J., Holtrop, G., Korpela, K., Duncan, S. H., Date, P., Farquharson, F., Johnstone, A. M., Lobley, G. E., et al. (2014). Impact of diet and individual variation on intestinal microbiota composition and fermentation products in obese men. *The ISME journal*, 8(11):2218–2230.

[Sansonetti, 2011] Sansonetti, P. (2011). To be or not to be a pathogen: that is the mucosally relevant question. *Mucosal immunology*, 4(1):8–14.

[Satokari et al., 2014] Satokari, R., Fuentes, S., Mattila, E., Jalanka, J., de Vos, W. M., and Arkkila, P. (2014). Fecal transplantation treatment of antibiotic-induced, noninfectious colitis and long-term microbiota follow-up. *Case reports in medicine*, 2014.

[Shastry and Sanjay, 2020] Shastry, K. A. and Sanjay, H. (2020). Machine learning for bioinformatics. In *Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications*, pages 25–39. Springer.

[Shin et al., 2015] Shin, N.-R., Whon, T. W., and Bae, J.-W. (2015). Proteobacteria: microbial signature of dysbiosis in gut microbiota. *Trends in biotechnology*, 33(9):496–503.

[Spor et al., 2011] Spor, A., Koren, O., and Ley, R. (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Reviews Microbiology*, 9(4):279–290.

[Tan et al., 2018] Tan, H., Yu, Z., Wang, C., Zhang, Q., Zhao, J., Zhang, H., Zhai, Q., and Chen, W. (2018). Pilot safety evaluation of a novel strain of bacteroides ovatus. *Frontiers in genetics*, 9:539.

[Tortora et al., 2007] Tortora, G. J., Funke, B. R., and Case, C. L. (2007). *Introducción a la microbiología*. Ed. Médica Panamericana.

[Tropini et al., 2017] Tropini, C., Earle, K. A., Huang, K. C., and Sonnenburg, J. L. (2017). The gut microbiome: connecting spatial organization to function. *Cell host & microbe*, 21(4):433–442.

[Turnbaugh et al., 2009] Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., et al. (2009). A core gut microbiome in obese and lean twins. *nature*, 457(7228):480–484.

[Verdam et al., 2013] Verdam, F. J., Fuentes, S., de Jonge, C., Zoetendal, E. G., Erbil, R., Greve, J. W., Buurman, W. A., de Vos, W. M., and Rensen, S. S. (2013). Human intestinal microbiota composition is associated with local and systemic inflammation in obesity. *Obesity*, 21(12):E607–E615.

[Wehrens and Kruisselbrink, 2018] Wehrens, R. and Kruisselbrink, J. (2018). Flexible self-organizing maps in kohonen 3.0. *Journal of Statistical Software*, 87(7):1–18.

[Wickham, 2016] Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

[Wickham et al., 2019] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

[World Health Organization, 2021] World Health Organization (2021). Body mass index - bmi. `https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi`, Last accessed on 2021-11-08.

[Wu et al., 2011] Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108.

[Xu et al., 2019] Xu, C., Zhu, H., and Qiu, P. (2019). Aging progression of human gut microbiota. *BMC microbiology*, 19(1):1–10.

[Yatsunenko et al., 2012] Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., et al. (2012). Human gut microbiome viewed across age and geography. *nature*, 486(7402):222–227.

[Yugi et al., 2016] Yugi, K., Kubota, H., Hatano, A., and Kuroda, S. (2016). Trans-omics: how to reconstruct biochemical networks across multiple 'omic'layers. *Trends in biotechnology*, 34(4):276–290.

Apéndice A

Artículo en ICAI 2021

Parte de los resultados de esta tesis fueron presentados en la *4th International Conference on Applied Informatics* (ICAI2021) llevada a cabo del 28 al 30 de Octubre del 2021 en Buenos Aires, Argentina. El trabajo fue seleccionado para su publicación [Burgos et al., 2021].

A Complementary Approach in the Analysis of the Human Gut Microbiome Applying Self-organizing Maps and Random Forest

Valeria Burgos[1( )], Tamara Piñero[1], María Laura Fernández[2], and Marcelo Risk[1]

[1] Instituto de Medicina Traslacional e Ingeniería Biomédica (IMTIB) - CONICET, Hospital Italiano de Buenos Aires, Instituto Universitario del Hospital Italiano, Buenos Aires, Argentina
valeria.burgos@hospitalitaliano.org.ar

[2] CONICET - Universidad de Buenos Aires, Instituto de Física del Plasma (INFIP), Buenos Aires, Argentina

Abstract. The human gastrointestinal tract is colonized by millions of microorganisms that make up the so-called gut microbiota, with a vital role in the well-being, health maintenance as well as the appearance of several diseases in the human host. A data mining analysis approach was applied on a set of gut microbiota data from healthy individuals. We used two machine learning methods to identify biomedically relevant relationships between demographic and biomedical variables of the subjects and patterns of abundance of bacteria. The study was carried out focusing on the two most abundant human gut microbiota groups, *Bacteroidetes* and *Firmicutes*. Both subsets of bacterial abundances together with the metadata variables were subjected to an exploratory analysis, using self-organizing maps that integrate multivariate information through different component planes. Finally, to evaluate the relevance of the variables on the biological diversity of the microbial communities, an ensemble-based method such as random forest was used. Results showed that age and body mass index were among the most important features at explaining bacteria diversity. Interestingly, several bacteria species known to be associated to diet and obesity were identified as relevant features as well. In the topological analysis of self-organizing maps, we identified certain groups of nodes with similarities in subject metadata and gut bacteria. We conclude that our results represent a preliminary approach that could be considered, in future studies, as a potential complement in health reports so as to help health professionals personalize patient treatment or support decision making.

Keywords: Gut microbiome · Self-organizing maps · Random forest · Precision medicine · Bioinformatics

H. Florez and M. F. Pollo-Cattaneo (Eds.): ICAI 2021, CCIS 1455, pp. 97–110, 2021.
https://doi.org/10.1007/978-3-030-89654-6_8

1 Introduction

Data science comprises different scientific fields of knowledge to target the analysis of complex and massive data. In particular, the increased interest on the application of machine learning algorithms to extract hidden associations or patterns in electronic health records, processing of medical images, prediction of a health situation or classification of patients has demonstrated the need for machine learning tools for reliable decision-making in healthcare and handling of biological data. The human gastrointestinal tract harbors millions of microorganisms which includes bacteria, archaea, fungi and viruses, interacting in symbiotic relationships between the host and each microbial community. This is known as the gut microbiota, while the collective genome of all symbiotic and pathogenic microorgnisms represents the gut microbiome. The establishment of a large part of the component communities that will remain in the adult life occurs at birth and during the first years of life [1,2] and its composition is shaped not only by the host genetics but also by environmental factors, nutritional status, age and lifestyle. Importantly, the gut microbiome plays an essential role in a number of health-beneficial functions (digestion, synthesis of essential vitamins and amino acids, absorption of calcium, magnesium and iron, fermentation of indigestible components, protection against pathogens, etc.) [3].

The rates of growth and survival of its component populations may fluctuate in response to temporary stressors, such as changes in diet or the consumption of antibiotics [4]. This potential for dynamic restructuring involves two important characteristics of the gut microbiome: plasticity and resilience [5,6]. Ongoing research in human and animal models highlights the importance of a healthy gut microbiome since persistent disbalances in composition and stability, known as dysbiosis, are associated to the onset and progression of chronic diseases that include obesity, irritable bowel syndrome, diabetes, cancer, and neurological diseases such as Parkinson's, among others [7].

The generation of biological knowledge from the large flow of data generated by new technologies in biomedical sciences has accelerated their transformation into data-centered fields. Thus, the study of the human gut microbiome represents a major challenge since it requires an interdisciplinary work between computer science and medicine. The interaction between these two fields will help obtain knowledge about gut bacteria interactions in human health and disease.

In the present work, we analyzed microbiome abundance data and the associated metadata using a machine learning approach: we used the visualization capabilities offered by self-organizing maps to identify patterns of multivariate data stored in multiple layers and additionally, we applied random forest to model the prediction of microbial diversity. For each analysis, we focused on the abundance levels of two major groups of the human gut bacteria, such as *Bacteroidetes* and *Firmicutes*.

2 Methods

2.1 Dataset

Microarray profiling data of human gut microbiota and anonymized metadata were obtained from the Dryad Digital Repository, as described by [8]. Briefly, the data matrix contained 1172 intestinal samples of western adults. In each sample, bacterial abundances were quantified using the HITChip phylogenetic microarray. This technology allows the assessment of relative abundances of gut bacteria through signal intensities of the targeted 16S rRNA gene, frequently used for the identification of poorly described or non cultured bacteria. Data contained hybridization signals for 130 genus-like phylogenetic groups. Subject metadata included age, sex, nationality, probe-level Shannon diversity, BMI group and subjectID. Geographical origin of the study subjects were: Central Europe (Belgium, Denmark, Germany, the Netherlands), Eastern Europe (Poland), Scandinavia (Finland, Norway, Sweden), Southern Europe (France, Italy, Serbia, Spain), United Kingdom/Ireland (UK, Ireland) and the United States (US). We used VIM and tidyverse R packages [9,10] to check for the presence of missing values (NAs). Records containing NAs were carefully removed without causing bias in the dataset. During the cleaning process, the category 'Eastern Europe' was turned out since it was represented by only one complete case. The final dataset to be used was represented by 1056 complete patient records containing 130 bacterial abundance data and subject metadata.

2.2 Self-organizing Maps (SOMs)

Self-organizing maps (also known as Kohonen maps) represent an optimal option to organize multidimensional data in a two-dimensional space by using a neural network. SOM uses the vector space as a model to represent data in a two-dimensional lattice: each value through N samples could be referred to as a data point in an N-dimensional space. Thousands of data points would therefore form data clouds in space, with a intrinsic topology due to geometric relationships. From this it follows that the greater the similarity in the data value level, the closer is the geometric space they occupy. To visualize the trained SOM, we used heatmaps for each variable to plot the degree of connectivity between adjacent output neurons through the use of a color intensity panel. In the case of multivariate datasets, the visualization of different heatmaps allows an overall analysis of the relations between the variables since maps are linked to each other by position: in each map, a node in a given location corresponds to the same unit in another map. The SOM map can be implemented in different topologies. Data were divided into two subsets by major groups of gut bacteria (*Bacteroidetes* and *Firmicutes*). We used a regular hexagonal 2D grid consisting of 750 neurons, in 30×25 grids. Data was logarithmically-scaled before training. We used the kohonen R package, which provides a standardized framework for SOMs.

2.3 Random Forest

Random forest (RF) is an ensemble learning method that can solve regression and classification problems. The algorithm uses a random subset of the training samples for each tree and a random subset of predictors in each step during the training process. These two sources of randomization make the algorithm robust to correlated predictors and more reliable at obtaining average outputs into a model. Data were divided into two subsets by major groups of gut bacteria (*Bacteroidetes* and *Firmicutes*). Bacterial abundance data and metadata were used as RF regressors to generate a diversity prediction model, which was performed using the RF regression algorithm provided by the R interface for h2o [11]. We used 10-fold cross validation for training the regression models and their performance were evaluated using Mean Absolute Error (MAE) as the error metric. After parameter tuning (mainly focused on the number of trees) through cross validation, the best RF regression model was selected.

3 Results

3.1 Characteristics of the Study Population

The degree of obesity is a relevant aspect in gut microbiome studies in terms of its influence on the microbiota composition [12,13]. This parameter, that can be obtained through the body mass index (BMI), indicates the nutritional status of an individual. Descriptive analysis of the study population showed that lean individuals were homogeneously distributed in all age groups, while overweight, obese and severe obese categories were more abundant in 45–60 year-old individuals. A large proportion of the underweight population was represented between 20–30 years old (Fig. 1).

The distribution of the different BMI categories in each geographic region showed that lean individuals represented approximately half of the proportions for all locations. For Scandinavia, Southern Europe, UK/Ireland (UKIE) and the US, the following proportion was represented by overweight subjects. In contrast, in Central Europe, the second proportion after lean individuals was represented by obese individuals. Morbid obese subjects were present only in Central Europe (Fig. 2).

3.2 SOM Analysis

Each component plane or map in a SOM represents one type of data: a two-dimensional lattice for each metadata variable (BMI, nationality, age, sex and diversity) as well as for each bacteria (whose relative abundance is represented in expression levels of the 16S rRNA gene). Since each map preserves shape and density, exploration of the geometric relationships between nodes allows a direct identification of similarities and differences between the layers.

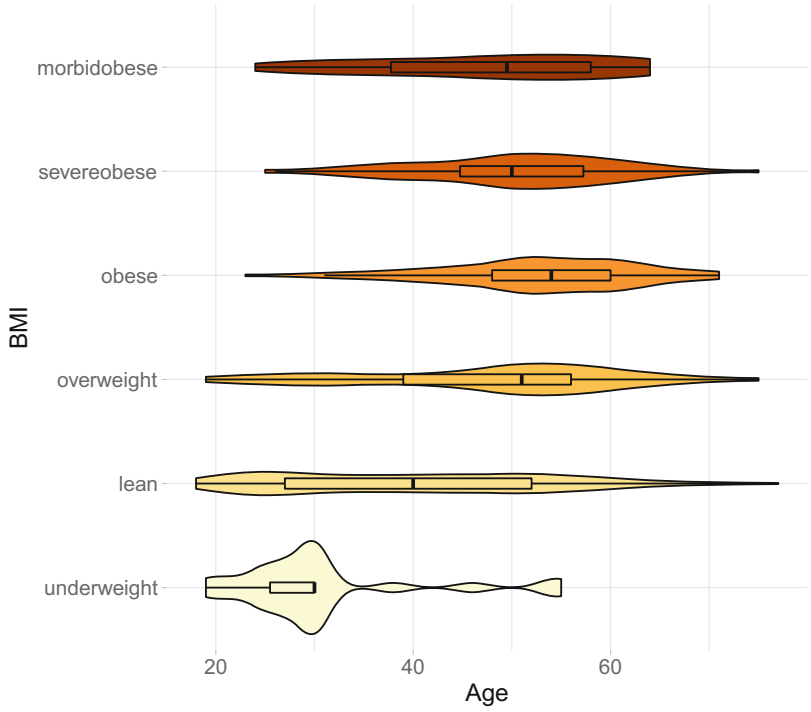

Fig. 1. Distribution of the BMI categories across age intervals in the study population.

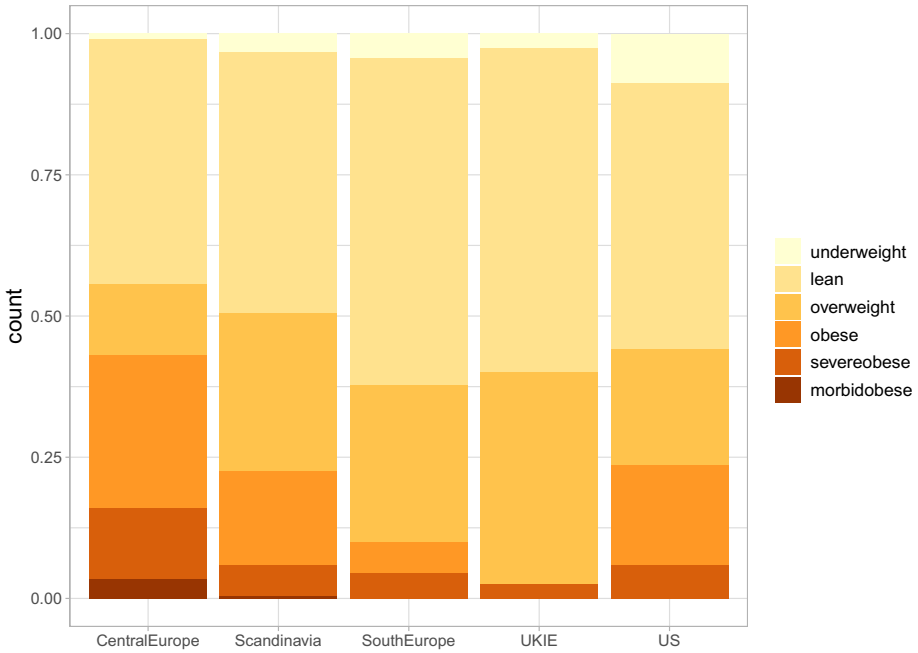

Fig. 2. Proportion of each BMI category of the study subjects across different geographic locations.

Bacteroidetes. After running the SOM algorithm using the *Bacteroidetes* subset, the different regions in each map indicated that the age distributed into two well-defined subregions of nodes for younger ages (around 20 years old), quite separated from a small group of nodes representing older individuals (older than 65 years old). Nodes representing 40–50 year old subjects were scattered throughout the map (Fig. 3A). Men and women were clearly distributed in two large, separated areas in the map (Fig. 3B). Scandinavian individuals mapped in a few groups of isolated nodes while Central European subjects were homogeneously distributed. Interestingly, the map identified an isolated group of nodes corresponding to individuals from the United States (Fig. 3C). Additionally, underweight and lean subjects mapped in two wide groups, while severe and morbid obese individuals mapped mainly in a small and defined group of nodes (Fig. 3E). The distribution of microbial diversity showed that two small and defined groups of nodes mapped low diversity values while higher values distributed into larger and clearly defined subsets of nodes across the map (Fig. 3D).

After SOM training, the abundance levels of each bacteria species of the *Bacteroidetes* phylum was also represented in a map. In the present dataset, several members belonging to this phylum were identified but no overlay between any bacterial map was observed (data not shown). However, since many members of the *Bacteroidetes* community have a relevant role in the host health, we chose to analyze two prominent bacteria whose relative abundances are known to be influenced by the host lifestyle and diet: a high fat and protein intake is associated with elevated microbial presence of *Bacteroides* species, while a high fiber intake is associated with high microbial levels of *Prevotella* species [14,15]. It appears that the abundances of these two *Bacteroidetes* members showed no overlay between any host metadata map because there are no coincidences in location (Fig. 3F and G).

The superimposition of the multiple maps described above allows to obtain some clear aspects of the data from the perspective of the *Bacteroidetes* subset:

- Low diversity values overlaps with a lower BMI (lean subjects) corresponding to young men from Central Europe and Scandinavia (indicated by a red circle in Fig. 3A–E).
- Interestingly, another subset of low diversity values overlaps with high BMI values (that is, severe to morbid obese) corresponding to middle-aged female individuals from Central Europe (indicated by a black circle in Fig. 3A–E).
- The small proportion of young subjects is equally represented in both men and women, while older individuals correspond exclusively to men.
- There are no women older than 65 years.
- US nationality corresponds to lean women in the range of medium values of microbial diversity.

Firmicutes. When subject metadata was trained using the *Firmicutes* subset, the different layers showed a more disperse variable distribution in the maps: middle-age individuals were homogeneously represented throughout the lattice while younger (around 20 years old) and older ages (older than 65 years) mapped

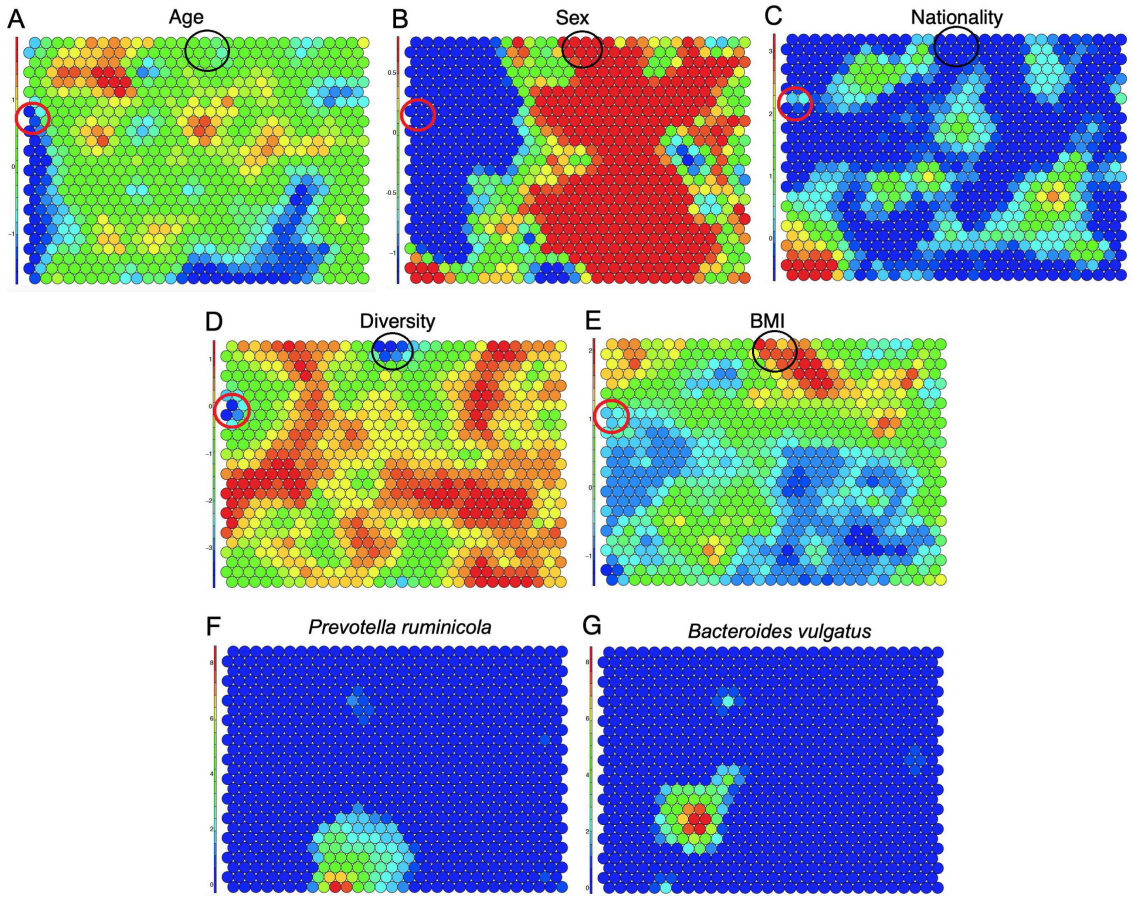

Fig. 3. SOM results for metadata variables associated with *Bacteroidetes*. The color index of each map is established based on the values for each variable. A) Age, scale adjusted for a range from 18 to 77 years old, blue = younger, red = older adults. B) Sex, blue = male, red = female. C) Nationality, color gradient beginning at Central Europe (level 0, color blue), Scandinavia (level 1, light blue), Southern Europe (level 2, green), UK/Ireland (level 3, orange) and the US (level 4, red). D) Diversity, scale adjusted for a range of 4.7 to 6.3 diversity index values beginning at blue (lowest diversity) to red (highest diversity). E) BMI, color gradient beginning at underweight (color blue), lean (light blue), overweight (green), obese (yellow), severe obese (orange) and morbid obese (red). In F) *P. ruminicola* and G) *B. vulgatus* the color gradient represents the level of abundance, blue = low, red = high. (Color figure online)

in small, discrete groups of nodes (Fig. 4A). Men and women were not as clearly separated in their node distribution as in the *Bacteroidetes* subset (Fig. 4B). Regarding nationality, a node pattern similar to *Bacteroidetes* was observed (Fig. 4C). High microbial diversity values predominated throughout the map while only three nodes mapped for low diversity values (Fig. 4D). Additionally, lean and underweight subjects predominated in most of the nodes and only a very small group of nodes grouped the highest BMI values, corresponding to the severe obese category (Fig. 4E).

The phylum *Firmicutes* is made up of around 250 different genera of bacteria, such as *Lactobacillus*, *Bacillus*, *Clostridium*, *Enterococcus*, and *Ruminicoccus*,

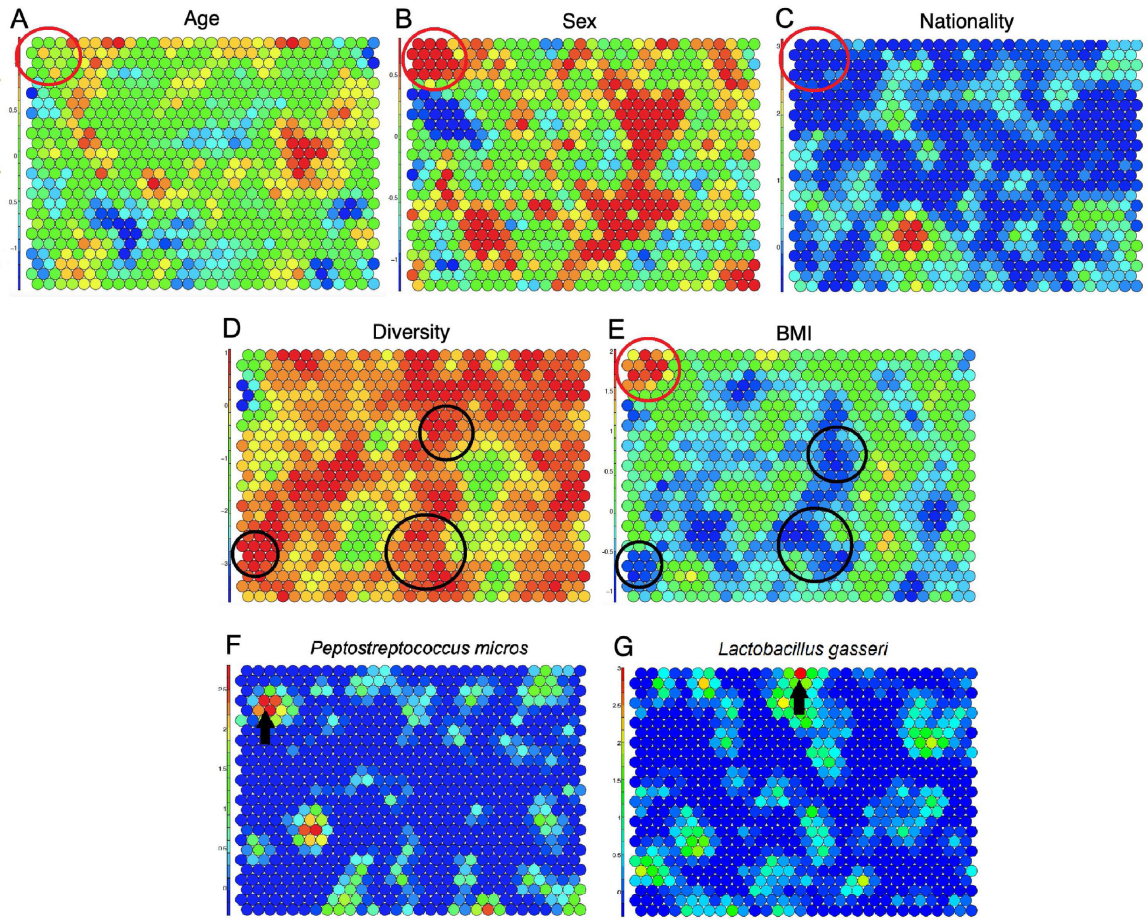

Fig. 4. SOM results for metadata variables associated with *Firmicutes*. The color index of each map is established based on the values for each variable. A) Age, scale adjusted for a range from 18 to 77 years old, blue = younger, red = older adults. B) Sex, blue = male, red = female. C) Nationality, color gradient beginning at Central Europe (level 0, color blue), Scandinavia (level 1, light blue), Southern Europe (level 2, green), UK/Ireland (level 3, orange) and the US (level 4, red). D) Diversity, scale adjusted for a range of 4.7 to 6.3 diversity index values beginning at blue (lowest diversity) to red (highest diversity). E) BMI, color gradient beginning at underweight (color blue), lean (light blue), overweight (green), obese (yellow), severe obese (orange) and morbid obese (red). In F) *P. micros* and G) *L. gasseri* the color gradient represents the level of abundance, blue = low, red = high. (Color figure online)

among other important members. In the present dataset, 74 members belonging to this phylym were identified and consequently, each generated a heatmap after SOM training (data not shown). However, superimposing multiple maps revealed that only one node corresponding to overweight individuals slightly coincided with higher values of abundance of a single species, *Peptostreptococcus micros* (indicated by an arrow in Fig. 4F). This represents an interesting result since several other members of *Firmicutes* have previously been associated to obesity [16,17]. Additionally, an overlay between high microbial diversity and higher values of abundance for *Lactobacillus gasseri* was observed (indicated by an arrow in Fig. 4G).

The map overlay between diversity and BMI categories allowed to observe that:

- Severe to morbid obese individuals were middle-age women from Central Europe (indicated by a red circle in Fig. 4A, B, C and E).
- Many of the maximum diversity values superimpose with the lowest BMI categories (indicated by a black circle in Fig. 4D and E).

3.3 Random Forest

Diversity is an important variable in microbiome research because it describes the richness (the number of classes) and distribution of the component microorganisms among classes. Understanding diversity in the intestinal microbial community also allows us to understand the impact of factors on bacteria distribution, such as the use of antibiotics, type of diet, degree of obesity, medical interventions and environmental factors, among others (reviewed in [18]). We developed a RF regression model of microbiome diversity. A very useful visual way to interpret RF results of the prediction model is through a ranking list of feature importance, which refers to the relative influence of each feature on the target variable. It considers whether a variable was selected to split and how much the squared error (over all trees) improved as a result.

Bacteroidetes. We observed that the age of the individuals was the most important factor in the regression model to predict the diversity of the gut microbiota. Regarding the bacterial abundances, it was observed that two species had the greatest relative importance in the identification of diversity: *Prevotella ruminicola* and *Bacteroides ovatus*. On the other hand, of the remaining metadata variables, the BMI of individuals had a remarkable position in the ranking of importance, followed by the geographic location. Gender was not important in the prediction of diversity. While the list of importance for each variable is informative, a better interpretation is obtained after scaling between 0 and 1 in a descending order of importance (Fig. 5A).

Firmicutes. When data was RF-trained considering this subset, the first three positions of the list were occupied by the BMI of the individuals, followed by Nationality and Age. Regarding the abundances of the 74 types of bacteria, it was observed that two members had the greatest relative importance in predicting diversity: *Lachnobacillus bovis* and *Granulicatella* (Fig. 5B).

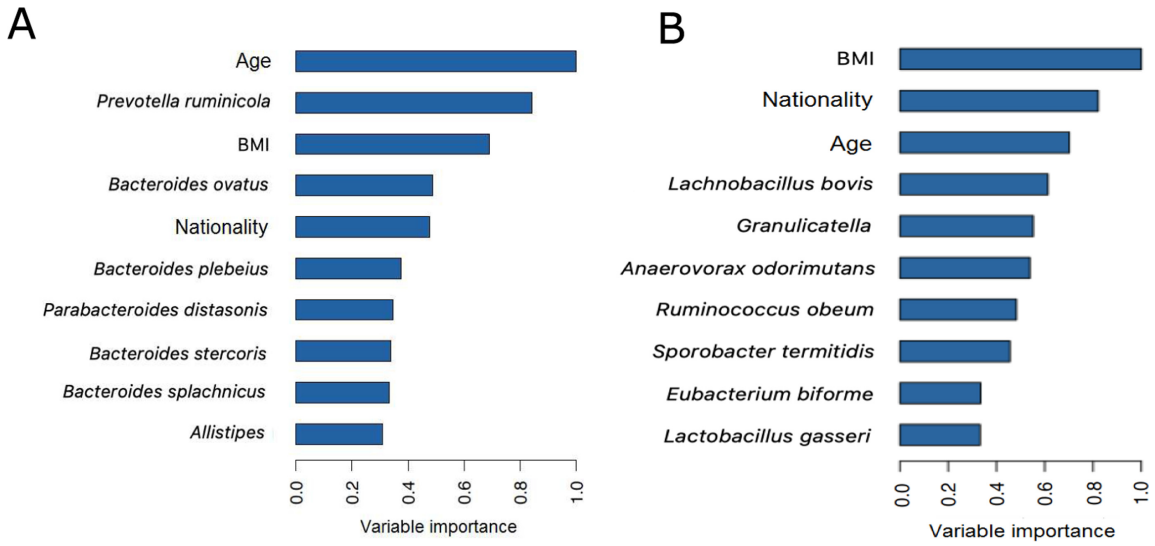

Fig. 5. Scale of the relative importance of the first ten variables on the diversity of *Bacteroidetes* (A) and *Firmicutes* (B) after implementing random forest.

4 Discussion

We sought to characterize the relations between subject metadata and specific bacterial members of the gut microbiome in a thousand western adults through the use of two machine learning methods that provide robust analytical visualizations, such as self-organizing maps and random forest. Bacteria of the human intestinal microbiome are taxonomically classified into six large groups or phyla, which in turn are subclassified into classes, orders, families, genera and species. The present work addressed the analysis on two of the most abundant phyla, *Bacteroidetes* and *Firmicutes*, which represent 90% of the gut microbiota [19].

The configuration of the SOM outcome maintains the topological structure of the input multidimensional data, in which similar values are mapped in the same or near node in a two dimensional map. Such topological preservation is of particular significance in the exploratory phase of omics data mining since there is generally no *a priori* knowledge of data structure. We presented the visualization of different superimposed heatmaps that allowed the exploration of relationships between input variables. This way of presenting SOM outcomes is similar to previous studies [20].

Our results showed that only one species of *Firmicutes*, *Peptostreptococcus micros*, was slightly associated to nodes that grouped overweight individuals, mostly middle-aged women. Notably, several previous studies have shown that *Peptostreptococcus micros* (later classified as *Parvimonas micra*) is one of various colorectal cancer (CRC) microbial markers [13,21]. This bacteria has also been reported to have a pathogenic role in periodontal diseases [22]. Considering the behavior of the oral microbiome during periodontal infection and its influence in health complications, such as diabetes, cardiovascular disease, and obesity [23], the significance of the results obtained here provides a basis for future studies on the possible role of gut bacteria as biological markers of a

developing overweight condition. Among bacteria with known beneficial roles, we observed that high abundance of *Lactobacillus gasseri* was related to nodes that grouped high diversity values. This result supports the previously reported role of *Lactobacillus gasseri* in the management of obesity and probiotic properties [24,25]. The topological structures of the metadata variables were slightly different depending whether *Bacteroidetes* or *Firmicutes* subsets were used for SOM training. For both age and sex, mapping distribution of the study subjects was more effective using the *Bacteroidetes* data. In the case of the different BMI categories, subject distribution was more effective using *Firmicutes* data.

Although the SOM results presented here allowed us to gain insight into the different regions of matching information underlying host metadata and the relative abundance of bacteria, we consider that a deeper approach of the use of SOM is needed, in terms of parameter configuration, such as size, dimensionality, shape, learning rate, among others. Considering that the relationships between gut microbiome and host BMI are dynamic and complex, self-organizing maps provide an excellent tool of visualization and dimensionality reduction that could serve as a complementary tool in a biomedical report.

Analysis of the microbiome diversity in the human body is essential to understand the structure, biology and ecology of its component communities. This analysis represents a critical first step in microbiome studies. When supervised learning through a regression random forest algorithm was used to determine which variables were important in the prediction of microbial diversity, we observed that both age and BMI category of the individuals appeared as the most relevant in the regression models generated for *Bacteroidetes* and *Firmicutes* subsets. The contribution of these physiological factors in shaping the gut microbiome has been reported previously: with age, the beneficial functions provided by a healthy gut microbiome begin to decrease in association to an increasing frequency of inflammatory processes and disease, especially in the elderly. Regarding the influence of BMI, various studies that compare the intestinal microbiota between obese and lean individuals indicate that the variation in the degree of diversity is associated with body weight (obese individuals present a low diversity, which means a higher BMI) [26–29].

Random forest regression also indicated that two *Bacteroidetes* species were the most relevant on diversity: *Prevotella ruminicola* is involved in the response of individuals to dietary supplements [30] and *Bacteroides ovatus* is a dominant species in the human intestine, previously identified as a next generation probiotic due to its preventive effects on intestinal inflammation (reviewed in [31]). On the other hand, two members of the *Firmicutes* group were identified as important at predicting diversity in this subset: *Lachnobacillus bovis* and *Granulicatella* whose relative abundances are reported to be influenced by the type of diet and the use of antibiotics in obese individuals [32,33]. Hence, some of the results obtained in both RF regression models are consistent with published microbiome research, indicating the robustness of the RF regression algorithm. A further analysis is needed that involves a complete parameter tuning so as to characterize the most accurate RF setting for a microbiome project.

In the last decade, the impulse provided by innovative developments in technology and the generation of large volumes of microbiome data has caused an increase in the use of machine learning methods in this field, such as microbial ecology, identification of certain bacteria to cancer and forensics, among others [34–36]. We consider that our study represents the start of a contribution to the vast field of microbiome research, although we need further refinements of the methodology used in order to validate the obtained models and improve performances.

The accumulating research of the gut microbiome and its influence on health and disease has accelerated the need for integration of multidisciplinary fields in its analysis. In general, health professionals (medical doctors, nurses, biochemists) are not prepared to work in all the steps along the data analysis process (cleaning, filtering, choice of algorithms, interpretation, etc.). Therefore, data science and machine learning can contribute to the translation of innovative results into valuable knowledge that provide decision support in microbiome-based precision medicine.

5 Conclusions

We used two robust computer science-based methods, such as self-organizing maps and random forest, to study the relationships between gut microbiome data and host information. Our results represent a preliminary approach that could be considered, in future studies, as a potential complement in health reports so as to help health professionals to individualize patient treatment or support decision making. Additionally, this work contributes to the increasingly growing area of gut microbiome interactions on human health and disease. However, further studies using other machine learning algorithms to validate the results obtained here are required.

References

1. Dominguez-Bello, M.G., et al.: Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc. Natl. Acad. Sci. **107**(26), 11971–11975 (2010)
2. Koenig, J.E., et al.: Succession of microbial consortia in the developing infant gut microbiome. Proc. Natl. Acad. Sci. **108**(Supplement 1), 4578–4585 (2011)
3. Rowland, I., et al.: Gut microbiota functions: metabolism of nutrients and other food components. Eur. J. Nutr. **57**(1), 1–24 (2017). https://doi.org/10.1007/s00394-017-1445-8
4. Kiewiet, M., et al.: Flexibility of gut microbiota in ageing individuals during dietary fiber long-chain inulin intake. Molecular Nutrition & Food Research, p. 2000390 (2020)
5. Ruggles, K.V., et al.: Changes in the gut microbiota of urban subjects during an immersion in the traditional diet and lifestyle of a rainforest village. Msphere, vol. 3(4) (2018)
6. Liu, H., et al.: Resilience of human gut microbial communities for the long stay with multiple dietary shifts. Gut **68**(12), 2254–2255 (2019)

7. Floch, M.H., Ringel, Y., Walker, W.A.: The microbiota in gastrointestinal pathophysiology: implications for human health, prebiotics, probiotics, and dysbiosis. Academic Press (2016)
8. Lahti, L., Salojärvi, J., Salonen, A., Scheffer, M., De Vos, W.M.: Tipping elements in the human intestinal ecosystem. Nat. Commun. **5**(1), 1–10 (2014)
9. Kowarik, A., Templ, M.: Imputation with the R package VIM. J. Stat. Softw. **74**(7), 1–16 (2016). https://doi.org/10.18637/jss.v074.i07
10. Wickham, H., et al.: Welcome to the tidyverse. J. Open Source Softw. **4**(43), 1686 (2019). https://doi.org/10.21105/joss.01686, http://dx.doi.org/10.21105/joss.01686
11. LeDell, E., et al.: h2o: R interface for 'h2o'. R package version 3(0.2) (2018)
12. Haro, C., et al.: Intestinal microbiota is influenced by gender and body mass index. PLoS ONE **11**(5), e0154090 (2016)
13. Yu, J., et al.: Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut **66**(1), 70–78 (2017)
14. De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J.B., Massart, S., Collini, S., Pieraccini, G., Lionetti, P.: Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc. Natl. Acad. Sci. **107**(33), 14691–14696 (2010)
15. Wu, G.D., et al.: Linking long-term dietary patterns with gut microbial enterotypes. Science **334**(6052), 105–108 (2011)
16. Koliada, A., et al.: Association between body mass index and firmicutes/bacteroidetes ratio in an adult ukrainian population. BMC Microbiol. **17**(1), 120 (2017)
17. Crovesy, L., Masterson, D., Rosado, E.L.: Profile of the gut microbiota of adults with obesity: a systematic review. Eur. J. Clin. Nutr. **74**(9), 1251–1262 (2020)
18. Reese, A.T., Dunn, R.R.: Drivers of microbiome biodiversity: a review of general rules, feces, and ignorance. MBio **9**(4), e01294-18 (2018)
19. Arumugam, M., et al.: Enterotypes of the human gut microbiome. Nature **473**(7346), 174–180 (2011)
20. Qian, J., et al.: Introducing self-organized maps (som) as a visualization tool for materials research and education. Results Mater. **4**, 100020 (2019)
21. Xu, J., et al.: Alteration of the abundance of parvimonas micra in the gut along the adenoma-carcinoma sequence. Oncol. Lett. **20**(4), 1 (2020)
22. Nagarajan, M., Prabhu, V.R., Kamalakkannan, R.: Metagenomics: implications in oral health and disease. In: Metagenomics, pp. 179–195. Elsevier (2018)
23. Goodson, J., Groppo, D., Halem, S., Carpino, E.: Is obesity an oral bacterial disease? J. Dent. Res. **88**(6), 519–523 (2009)
24. Selle, K., Klaenhammer, T.R.: Genomic and phenotypic evidence for probiotic influences of lactobacillus gasseri on human health. FEMS Microbiol. Rev. **37**(6), 915–935 (2013)
25. Mahboubi, M.: Lactobacillus gasseri as a functional food and its role in obesity. Int. J. Med. Rev. **6**(2), 59–64 (2019)
26. Turnbaugh, P.J., et al.: A core gut microbiome in obese and lean twins. Nature **457**(7228), 480–484 (2009)
27. Yatsunenko, T., et al.: Human gut microbiome viewed across age and geography. Nature **486**(7402), 222–227 (2012)
28. Dominianni, C., et al.: Sex, body mass index, and dietary fiber intake influence the human gut microbiome. PLoS ONE **10**(4), e0124599 (2015)
29. Bosco, N., Noti, M.: The aging gut microbiome and its impact on host immunity. Genes & Immunity, pp. 1–15 (2021)

30. Chung, W.S.F., et al.: Relative abundance of the prevotella genus within the human gut microbiota of elderly volunteers determines the inter-individual responses to dietary supplementation with wheat bran arabinoxylan-oligosaccharides. BMC Microbiol. **20**(1), 1–14 (2020)
31. Tan, H., et al.: Pilot safety evaluation of a novel strain of bacteroides ovatus. Front. Genet. **9**, 539 (2018)
32. Salonen, A., et al.: Impact of diet and individual variation on intestinal microbiota composition and fermentation products in obese men. ISME J. **8**(11), 2218–2230 (2014)
33. Reijnders, D., et al.: Effects of gut microbiota manipulation by antibiotics on host metabolism in obese humans: a randomized double-blind placebo-controlled trial. Cell Metab. **24**(1), 63–74 (2016)
34. Ai, D., Pan, H., Han, R., Li, X., Liu, G., Xia, L.C.: Using decision tree aggregation with random forest model to identify gut microbes associated with colorectal cancer. Genes **10**(2), 112 (2019)
35. Thompson, J., Johansen, R., Dunbar, J., Munsky, B.: Machine learning to predict microbial community functions: an analysis of dissolved organic carbon from litter decomposition. PLoS ONE **14**(7), e0215502 (2019)
36. Topçuoğlu, B.D., Lesniak, N.A., Ruffin, M.T., IV., Wiens, J., Schloss, P.D.: A framework for effective application of machine learning to microbiome-based classification problems. MBio **11**(3), e00434-20 (2020)