UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

MÉTODOS DE APRENDIZAJE DE MÁQUINA PARA EL ESTUDIO DE ASOCIACIONES MICROBIOMA-ENFERMEDADES

TESIS

QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN CIENCIAS

PRESENTA:
DANIEL NERI ROSARIO

TUTOR PRINCIPAL
Dr. OSBALDO RESENDIS ANTONIO
INMEGEN

MIEMBROS DEL COMITÉ TUTOR

Dra. BLANCA ITZEL TABOADA RAMÍREZ
IBT

Dr. LUIS DAVID ALCARAZ PERAZA
Facultad de Ciencias

Ciudad de México. Junio, 2022

Oficio de asignación de jurado

CGEP/PMDCBQ/48
FCC355-630109E686FEB
Asunto: Jurado de examen

SINODALES DESIGNADOS
Presente

Estimado académico:

Los miembros del Subcomité Académico en reunión ordinaria del **15 de agosto de 2022**, conocieron la solicitud de asignación de **JURADO DE EXAMEN** para optar por el grado de **Maestro en Ciencias** del/la estudiante **NERI ROSARIO DANIEL**, con la tesis **"Métodos de Aprendizaje de Máquina para el estudio de asociaciones Microbioma- Enfermedades"**, dirigida por el/la Dr(a). **RESENDIS ANTONIO OSBALDO**.

De su análisis se acordó nombrar el siguiente jurado en el que se encuentra usted incluido:

			ACEPTA	FECHA	FIRMA
SEGOVIA FORCELLA LORENZO PATRICK	PMDCBQ	PRESIDENTE	SI X NO___	22 / 09 / 22	
MENDOZA SIERRA LUIS ANTONIO	PMDCBQ	SECRETARIO	SI X NO___	30 / 09 / 22	
ARCINIEGA CASTRO MARCELINO	PMDCBQ	VOCAL	SI X NO___	22 / 09 / 22	
DOMÍNGUEZ HÜTTINGER ELISA	PMDCBQ	VOCAL	SI X NO___	23 / 09 / 22	
GUTIÉRREZ RÍOS ROSA MARÍA	PMDCBQ	VOCAL	SI X NO___	22 / 09 / 22	

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

A t e n t a m e n t e
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Cd. Universitaria, Cd. Mx., a 22 de agosto de 2022

Coordinadora
Dra. Claudia Lydia Treviño Santa Cruz

Dedicatorias

A la UNAM por haberme otorgado las herramientas para mi desarrollo personal y académico, en especial al programa de Ciencias Bioquímicas que durante esta etapa me permitieron desarrollarme en el ambiente de investigación.

A Paulina S. por compartir estos momentos con escucha, amor y cariño.

A mis padres y a mi hermana, son el ejemplo y el soporte en mi vida.

Al Dr. Osbaldo, por ser un maestro y guía en este proceso.

A mis tutores, la Dra. Taboada y el Dr. Alcaraz, sus consejos sirvieron para dirigir este proyecto.

A mis compañeros de laboratorio, por su amistad y apoyo.

Agradezco al Consejo Nacional de Ciencia y Tecnología (CONACyT) (CVU: 1083211) y Coordinación General de Estudios de Posgrado (CGEP) por el apoyo otorgado durante estos dos años para realizar mis estudios de maestría, mediante su programa de becas.

Tabla de acrónimos.

Abreviatura	Significado
AB	Ácidos biliares
ACV	Accidente cerebrovascular
ADH	Hormona antidiurética (del inglés, *Anti-diuretic hormone*)
AGCC	Ácidos grasos de cadena corta
APRIL	Ligando inductor de la proliferación (del inglés, *A proliferation-inducing ligand*)
ASV	Variantes de secuencia de amplicones (ASV, del inglés *Amplicon Sequence Variant*)
AUC	Área bajo la curva (AUC del inglés, *Area Under a Curve*)
BAFF	Factor de activación célula B (del inglés, *B cell–activating factor*)
C	Clasificación
CRC	Cáncer colorrectal
CV	Validación cruzada (CV, del inglés *cross validation*)
DL	Aprendizaje profundo (del inglés, *Deep learning*)
DMT1	Diabetes mellitus tipo 2
DMT2	Diabetes mellitus tipo 2
IC	Intervalo de confianza
IDF	Federación internacional de diabetes (del inglés, *International Diabetes Federation*)
INEGI	Instituto Nacional de Estadística y Geografía
EAP	Enfermedad arterial periférica
EII	Enfermedad inflamatoria intestinal
GLP-1	Péptido similar al glucagón 1 (del inglés, *Glucagon-like peptide-1*)
GLUT-4	Transportador de glucosa tipo 4 (del inglés, *Glucose transporter type 4*)

GPR	Receptor acoplado a proteína G (del inglés, *G protein-coupled receptor*)
HAS	Hipertensión arterial sistémica
Hb	Hemoglobina
IAM	Infarto agudo al miocardio
IC	Intervalo de confianza
IgA	Inmunoglobulina A
IL	Interleucina
IRS	Sustrato del receptor de la insulina (del inglés, *Insulin receptor substrate*)
LPS	Lipopolisacáridos
MALT	Tejido linfoide asociado a mucosa (del inglés, *Mucosa-associated lymphoid tissue*)
MCP-1	Proteína quimiotáctica de monocitos tipo 1 (del inglés, *Monocyte chemoattractant protein-1*)
ML	Aprendizaje de máquina (del inglés, *Machine learning*)
NF-κB	Factor nuclear κB (del inglés, *Nuclear factor κB*)
SHAP	Explicaciones aditivas de valores Shapley (del inglés, *SHapley Additive Explanations*)
SGLT	Cotransportadores sodio-glucosa tipo (del inglés, *Sodium-glucose cotransporter*)
TLR	Receptor tipo toll (del inglés, *Toll like receptor*)
TMA	Trimetilamina
TNF-α	Factor de necrosis tumoral-α (del inglés, *Tumor necrosis factor-α)*

Índice de figuras.

Abstract

A direct link between the gut microbiota (GM) and the progression of type 2 diabetes mellitus (T2D) in individuals has been described. We propose using supervised Machine Learning (ML) methods to identify predictive taxa for patients with prediabetes (pre-T2D) and T2D. For this, we obtained the GM profile (16s rRNA) in a cohort of 410 Mexican naïve patients, stratified into normoglycemic (n= 213), pre-T2D (n= 150), and T2D (n= 47) individuals. Using the abundances and taxonomies of the GM of these individuals, we performed three studies in silico to identify the bacteria associated with the progression of the condition. The first study, labeled classification 1 (C-1), includes a comparison study between normoglycemic subjects vs. T2DM patients. On the other hand, in classification 2 (C-2), we proceeded to identify bacteria that differentially classified normoglycemic subjects vs. pre-T2DM patients. Finally, in classification 3 (C-3), we identified those bacteria that allow classifying a multiclass group composed of: normoglycemic subjects, patients with T2DM, and patients with pre-T2DM. We used six different algorithms in each classification, including Logistic regression, naïve Baye, Decision-tree, Random Forest, XGBoost, and Multilayer perceptron. Random Forest obtained the best predictive performance to classify T2D patients (AUC: 0.98) and pre-T2D patients (AUC: 0.91). During multiclass classification, we obtain an AUC of 0.95 using the XGBoost algorithm. Through an explanatory ML analysis approach, we identify a set of taxa including *Allisonella, Slackia, Ruminococus_2, Megaspgaera, Escherichia/Shigella,* and *Prevotella* for predicting patients with T2D compared to normoglycemic subjects. In addition, the most critical genera for predicting patients with pre-T2D compared to normoglycemic subjects were: *Anaerostipes, Intestinibacter, Prevotella_9, Granulicatella,* and *Veillonella*. According to the literature, these genera may play a role in the pathophysiology of the disease. These results contribute to exploring the relationship between GM and the development of biomarkers to accurately identify people at high risk for T2D Mexican patients, with the perspective of receiving preventive and personalized treatments.

1 RESUMEN.

Actualmente los algoritmos de aprendizaje de máquina (Machine Learning (ML) en inglés) son utilizados como herramientas para identificar biomarcadores en la microbiota asociados con el progreso de enfermedades. De forma notable, se ha descrito una asociación directa entre la microbiota intestinal y la progresión de los individuos con diabetes mellitus tipo 2 (DMT2). Bajo este contexto, se propone la utilización de métodos de ML supervisados que permitan identificar taxones predictivos para individuos con DMT2 o pre-DMT2. Por tal motivo, el trabajo presentado en esta tesis integra la caracterización del perfil del microbiota intestinal mediante secuenciación ARNr 16s en una cohorte de 410 pacientes mexicanos sin tratamiento previo, estratificados en individuos sanos (n= 213), prediabéticos (n= 150) y diabéticos (n= 47). Utilizando las abundancias y taxonomías de la microbiota intestinal de estos individuos, realizamos tres estudios *in silico* con el objetivo de identificar las bacterias asociadas al progreso del padecimiento. El primer estudio, etiquetado como la clasificación 1 (C-1), incluye un estudio de comparación entre pacientes sanos contra pacientes con DMT2. Por otra parte, en la clasificación 2 (C-2) se procedió a identificar bacterias que diferencialmente clasificaron pacientes sanos contra pacientes con pre-DMT2. Finalmente, en la clasificación 3 (C-3) realiza una clasificación multiclase considerando: individuos sanos, pacientes con pre-DMT2 y pacientes con DMT2. Con la finalidad de encontrar el clasificador con mayor desempeño, en cada clasificación se comparó el rendimiento predictivo entre los distintos métodos de ML lineales (Regresión Logística Binaria y Naive Bayes) y no-lineales (Árboles de Decisiones, Random Forest, XG Boost, Perceptrón multicapas). El modelo con mejor rendimiento predictivo fue seleccionado para realizar un análisis de interpretación *post-hoc* identificando los géneros bacterianos de mayor importancia para la correcta clasificación del fenotipo en DMT2 en población mexicana. Nuestro estudio nos permitió concluir que el mejor método para

predecir a pacientes diabéticos (C-1) y prediabéticos (C-2) fue Random Forest con un rendimiento predictivo alto (AUC: 0.97 con una desviación estándar (DE) 0.1; AUC: 0.8 con DE 0.05 en la C-2). En el caso de la clasificación multiclase (C-3) concluimos que XGBoost tuvo el mejor rendimiento predictivo con un Cohen de Kappa score de 0.62 y una DE 0.08. A partir de estos resultados, realizamos el análisis de interpretación del modelo para distinguir las bacterias asociadas a cada grupo. Este análisis nos llevó a concluir que Allisonella*, Slackia, Ruminococus_2, Megaspgaera, Escherichia/Shigella, y Prevotella* se encuentra entre los géneros más relevantes para predecir a pacientes con DMT2 comparado con individuos sanos. En cambio, los géneros más importantes para predecir a pacientes con pre-DMT2, comparado con pacientes sanos fueron: *Anaerostipes, Intestinibacter, Prevotella_9, Granulicatella, y Veillonella*. Consistentemente con la literatura, algunos de los géneros encontrados en este estudio han sido descritos previamente en estudios de microbiota intestinal en otras poblaciones y se ha planteado un posible rol en la fisiopatología de la enfermedad. Nuestro trabajo resalta la necesidad de realizar estudios de población específica para el desarrollo de tratamientos individualizados basados en la microbiota intestinal. Estos resultados centrados en nuestra población mexicana contribuyen a explorar la relación existente entre la microbiota intestinal y el desarrollo de biomarcadores para identificar con precisión a personas con alto riesgo de desarrollar diabetes, con la perspectiva de recibir tratamientos preventivos y personalizados.

2 INTRODUCCIÓN

Actualmente, el estudio de las interacciones huésped-microbioma ha tenido relevancia sustancial para mejorar la predicción y entendimiento de enfermedades, debido a la gran influencia que tiene la microbiota en el estado de salud del ser humano (Contreras et al. 2016). Con base en esta perspectiva, se han encontrado resultados prometedores en modelos de asociación de microbiota y fenotipo clínico, específicamente para padecimientos complejos como cáncer colorrectal (CRC), enfermedad inflamatoria intestinal (EII), cirrosis hepática, obesidad y diabetes mellitus tipo 2 (DMT2) (Zhou and Gallins 2019). Entre estos estudios, el caso de la DMT2 toma relevancia importante dada la frecuencia y costo poblacional a nivel mundial con una prevalencia cerca de 10.5% (alrededor de 44% de personas sin diagnosticar, mediante estimaciones de estudios basados en población) (Tönnies et al. 2021; Magliano, Boyko, and IDF Diabetes Atlas 10th edition scientific committee, n.d.). Esta epidemia ha ido en aumento de forma alarmante en la última década, en gran parte relacionada con las tendencias mundiales de obesidad y sedentarismo (Tönnies et al. 2021; Chatterjee, Khunti, and Davies 2017). En un esfuerzo de detener este crecimiento, se ha descrito una relación directa entre los individuos con DMT2 y la microbiota intestinal, estando implicada en el inicio y la progresión de la enfermedad.

Notablemente, estudios hechos en cohortes de población china, europea, y estadounidense se ha descrito una pérdida en la homeostasis de la microbiota intestinal (conocida como disbiosis intestinal) en los individuos con DMT2 (Qin et al. 2012; Karlsson et al. 2013; Sikalidis and Maykish 2020). La disbiosis se refiere a los cambios en la composición y función de la microbiota que se asocia a enfermedades humanas (Hooks and O'Malley 2017). Desde el punto de vista fisiopatológico, en pacientes con DMT2 la disbiosis intestinal se asocia con un aumento de la permeabilidad intestinal, estimulación de una inflamación sistémica de

bajo grado y a una inadecuada modulación del sistema inmunitario y metabolismo celular por parte de la microbiota que en conjunto afecta a el progreso de la enfermedad (Zhao et al. 2020).

Con la finalidad de buscar nuevas estrategias que contribuyan a cesar el creciente problema global que es la DMT2, se han realizado algunos esfuerzos para identificar la asociación entre patrones en la composición de la microbiota intestinal (taxones bacterianos) y los pacientes con DMT2. En este contexto, dichos estudios tienen el objetivo de desarrollar intervenciones diagnósticas y terapéuticas personalizadas en pacientes con DMT2 o pacientes con alto riesgo de desarrollar este padecimiento (conocido como pre-DMT2). Actualmente se pone especial atención en los países en vías de desarrollo, como México, ya que se ha descrito una alta mortalidad e incidencia de DMT2 en estas poblaciones (Contreras et al. 2016).

Es de relevancia mencionar que la asociación entre la microbiota intestinal y DMT2 varía dependiendo de diferentes variables sociodemográficas del individuo. Por ejemplo, un estudio en una cohorte de 345 pacientes chinos con DMT2 se describió que una característica diferencial respecto a individuos sanos es la disminución de las especies productoras de butirato, como *Roseburia intestinalis y Faecalibacterium prausnitzii* (Qin et al. 2012). Por el contrario, en un segundo estudio realizado en pacientes europeos con DMT2 se encontró una disbiosis en ciertas especies como *Lactobacillus gasseri, Streptococcus mutans y Clostridium clostridioforme* (Karlsson et al. 2013). Estos estudios plantean el hecho de que la asociación entre microbiota intestinal y DMT2 es específica de cada población y sujeta a factores externos como la dieta, antecedentes heredofamiliares, estilo de vida, uso de fármacos, entre otros (Hasan and Yang 2019).

Por otra parte, debido a la alta dimensionalidad y complejidad que representan los datos de alto rendimiento que son utilizados para caracterizar la microbiota, existe la necesidad de aplicar diferentes acercamientos computacionales que coadyuven a los métodos estadísticos tradicionales para estudiar la asociación microbioma-enfermedades. Esta labor no es trivial, y para resolver este problema, se han propuesto múltiples métodos de aprendizaje de máquina (ML, del inglés Machine Learning) que resuelvan tareas de clasificación (genéricamente llamados métodos supervisados). El objetivo en esta tarea es crear modelos de clasificación que identifiquen la asociación microbioma-enfermedad de forma precisa y robusta, con la capacidad de detectar las complejas interacciones y efectos no lineales entre las comunidades microbianas. Por otro lado, ha llamado la atención aplicar métodos de aprendizaje profundo (DL, del inglés Deep Learning), debido a los avances recientes que ha habido en el campo. Se tiene la expectativa que al usar algoritmos de DL se puedan desarrollar modelos con altos valores predictivos para identificar los distintos fenotipos del individuo, en el área de la investigación del microbioma (Borenstein and Muller 2021).

En este proyecto de investigación, proponemos aplicar diferentes métodos de ML para identificar los taxones clave que puedan ser capaces de predecir los fenotipos clínicos de una cohorte de pacientes mexicanos estratificados en grupos de pre-DMT2, DMT2, e individuos sanos. De esta manera, en nuestro estudio comparamos el rendimiento de seis distintos algoritmos de ML para clasificar el estado de salud frente al de enfermedad utilizando los datos del microbioma intestinal caracterizados por el gen marcador 16s ARNr (ARN ribosomal). De forma general, nuestro estudio aporta un análisis detallado de las posibles bacterias que pueden identificar y clasificar los distintos estadios de la DMT2 en pacientes mexicanos, incluyendo estadios de prediabetes hasta la forma declarada de este padecimiento. Además, mediante revisión de la literatura sobre el estudio de las funciones de estos taxones proponemos el papel que tienen estos en la patogenia de la enfermedad. Consideramos que los resultados obtenidos nos permitirán identificar bacterias cuyas

abundancias nos permitan monitorear el progreso de la enfermedad, y a largo plazo identificar posibles biomarcadores bacterianos para el progreso y tratamiento de la enfermedad en población mexicana.

3 MARCO TEÓRICO

3.1 Diabetes Mellitus tipo 2.

3.1.1 Definición.

La DMT2 es una de las principales causas de enfermedad y mortalidad en todo el mundo. La prevalencia a nivel mundial de la diabetes mellitus (DM) es del 10,8% en las mujeres y el 13,3% en los hombres. Actualmente se estima que 537 millones de personas en el mundo padecen DM y se estima que para el año 2030 se alcanzará una prevalencia de 643 millones, y para el año 2045 aumente a 783 millones, lo que significa un incremento de 43% en la prevalencia. Estos datos confirman que la DM es una de las emergencias sanitarias más importantes en atender para el siglo XXI (Tönnies et al. 2021; Chatterjee, Khunti, and Davies 2017).

El término DM hace referencia a las enfermedades crónicas caracterizadas por hiperglucemia, debido a un déficit en la producción de insulina o porque la insulina elaborada no puede ser utilizada de forma adecuada. La insulina es una hormona esencial producida en las células β del páncreas, su principal función ocurre al unirse a los receptores membranales en diversos tejidos (como hígado, músculo, tejido adiposo) y activar/inhibir diferentes señales metabólicas que permite la entrada de glucosa desde el torrente sanguíneo hacia dichos tejidos para utilizarse en procesos catabólicos y anabólicos en el interior de la célula. La insulina también es esencial, no solo para el metabolismo de carbohidratos, sino también para el metabolismo de las proteínas y los lípidos (Pessin and

Saltiel 2000). Por tal motivo, la deficiencia de la producción de insulina o la incapacidad de las células para responder a ella da lugar a niveles elevados de glucosa en sangre (hiperglucemia) y diferentes alteraciones metabólicas, como dislipidemias e hipertensión arterial sistémica (HAS), que conducen a largo plazo a complicaciones vasculares crónicas.

Hay principalmente dos tipos de diabetes, la diabetes mellitus tipo 1 (DMT1) y la 2 (DMT2). La DMT2 es el tipo de diabetes más común en adultos (>90%) y se caracteriza al inicio de la enfermedad por la incapacidad de las células del organismo para responder adecuadamente a la insulina, un fenómeno denominado resistencia a la insulina. En respuesta a la aparición de la resistencia a la insulina, se provoca un aumento compensatorio de la producción de insulina, presentando un aumento del tamaño (hipertrofia) y un aumento del número (hiperplasia) de las células β del páncreas (Koenig et al. 2019). Conforme progresa la DMT2 se da un agotamiento funcional de las células β del páncreas, debido a esta respuesta compensatoria exagerada, provocando una producción anormal de la hormona insulina. Aunado a lo anterior, hay ciertos factores de riesgo que se asocian a la resistencia a la insulina, como la obesidad, la falta de ejercicio, dieta y diversos factores poli-genéticos que abordaremos más adelante (Duncan 2006).

En relación con la DMT2 existe un estadio intermedio conocido como prediabetes o pre-DMT2. Este estadio engloba a las personas con alteración en la tolerancia en la glucosa oral y/o alteración en la glucosa en ayunas que no sobrepasan el umbral de los criterios diagnósticos para DMT2 (American Diabetes Association 2013). Es fundamental su reconocimiento ya que estos individuos tienen mayor riesgo de desarrollar DMT2 y otras enfermedades incluyendo dislipidemias, hipertensión y un aumentado riesgo cardiovascular (Cai et al. 2020). Estudios previos sugieren que alrededor del 15-25% de los pacientes con pre-DMT2 en los primeros 3-5 años, y eventualmente alrededor del 70% de los pacientes con pre-DMT2 van a progresar a DMT2 en algún punto de su vida (Hostalek 2019). Suele

haber un largo período de este estado intermedio de pre-DMT2, ya que el momento exacto de la aparición de la DMT2 suele ser muy difícil de determinar (Hostalek 2019). Es importante seguir estudiando a los pacientes con pre-DMT2, ya que un tratamiento temprano durante la fase asintomática mejora el resultado a largo plazo y puede disminuir el riesgo de convertirse en individuos con DMT2 (Gong et al. 2021).

Varios estudios clínicos señalan que la DMT2 puede ser prevenida o retardada, y cada vez hay más evidencia que en ciertas condiciones es posible una remisión de la DMT2, como en pacientes tratados con cirugía bariátrica (Schauer, Hanipah, and Rubino 2017). Esto contrasta con la DMT1, que no puede ser prevenida y ocurre en menos del 10% de todos los casos de diabetes, siendo el principal tipo de diabetes en la infancia (Bullock and Sheff 2017). La DMT1 es ocasionada por una pérdida en la tolerancia del sistema inmunológico causando una respuesta inmune celular (principalmente linfocitos T cooperadores CD4+) contra diversos antígenos en las células β del páncreas, ocasionando una falta de insulina endógena necesaria para mantener los niveles normales de glucosa en sangre. La causa exacta de esta pérdida de la tolerancia contra antígenos propios (conocida como autoinmunidad) se desconoce, sin embargo, se asocia a una susceptibilidad genética. Los haplotipos mejor descritos son HLA-DR3, HLA-DR4, y además se han asociado distintos factores ambientales como infecciones virales (Krzewska and Ben-Skowronek 2016). Por tal motivo las personas con DMT1 requieren insulina exógena para sobrevivir y mantener niveles adecuados de glucosa con el objetivo de retrasar la progresión de la enfermedad.

Los pacientes con DMT2 pueden presentar síntomas similares a los de la DMT1, pero en general, los síntomas son mucho menos dramáticos y la enfermedad puede ser completamente asintomática. Lo anterior resalta la importancia de los métodos de escrutinio para identificar a los individuos con diabetes. Los síntomas clásicos de la hiperglucemia son: poliuria (aumento de la frecuencia miccional), polidipsia (aumento de la

sensación de sed), polifagia (aumento del apetito e ingesta calórica), y pérdida de peso. Estos datos clínicos generalmente se observan de forma retrospectiva, o son resultado de una complicación aguda de la enfermedad como la cetoacidosis diabética o estado hiperglucémico hiperosmolar que ponen en peligro la vida del individuo con diabetes (Koenig et al. 2019).

La poliuria se produce cuando la concentración de glucosa en sangre se eleva por encima de valores de 180 mg/dL (10 mmol/L), superando el umbral de los cotransportadores sodio-glucosa tipo 1 (SGLT1, por sus siglas en inglés *Sodium-glucose cotransporter type 1*) y SGLT2. Estas proteínas realizan de forma fisiológica la reabsorción de glucosa en los túbulos renales, y al superar su capacidad de absorción, provoca un aumento de la excreción de glucosa por la orina, llamada glucosuria. La glucosuria causa una diuresis osmótica en el paciente, originando un aumento de la excreción de agua y electrolitos por la orina, observado clínicamente como poliuria y signos de deshidratación. Esta depleción del agua corporal aumenta la secreción de la hormona antidiurética (ADH) o vasopresina que activa el centro de la sed a nivel del hipotálamo, es decir polidipsia (Meigs et al. 2003). Los pacientes al intentar compensar sus pérdidas de volumen con bebidas azucaradas concentradas agravan su hiperglucemia y la diuresis osmótica que ocasiona un aumento en su sintomatología (Watanabe et al. 2019).

Como mencionamos previamente, en los pacientes con DMT2, a pesar de que existen niveles elevados de glucosa en sangre esta no puede acceder a los tejidos respondedores de insulina, como el tejido adiposo y músculo esquelético. Por tal motivo, el tejido adiposo empieza a degradar lípidos (proceso llamado lipólisis) y el tejido muscular empieza a degradar proteínas (proceso llamado proteólisis) se manifiesta clínicamente como una pérdida de peso y un aumento en el apetito o polifagia (Kahn, Cooper, and Del Prato 2014).

Se recomienda realizar pruebas de niveles de glucosa en sangre a los pacientes con síntomas clásicos de DM, ya descritos previamente, y también realizar tamizaje a los pacientes asintomáticos con alto riesgo de prediabetes o diabetes (por ejemplo, pacientes con sobrepeso u obesidad y la presencia de otros factores de riesgo adicionales para desarrollar DMT2 como síndrome de ovario poliquístico, dislipidemia, familiar con primer grado con diabetes, antecedente de diabetes gestacional, entre otros) (Hussain et al. 2021). En México el cribado y diagnóstico de la diabetes se encuentran bajo los estándares internacionales de acuerdo a la Federación Internacional de Diabetes (IDF del inglés, *International Diabetes Federation*) y la Asociación Americana de Diabetes (ADA, del inglés *American Association of Diabetes*) (Atlas and Others 2015; Diabetes Association 2022). Se identifican como individuos con DM si tienen valores de glucosa en ayunas (más de 8 horas de ayuno) ≥ 126 mg/dl, también se diagnóstica si la glucosa en sangre a las dos horas después de consumir 75 gr de glucosa es ≥ 200 mg/dl y/o mediante hemoglobina (Hb) glucosilada por arriba ≥ 6.5% (Cosentino et al. 2020).

La DMT2 es un trastorno común, con una prevalencia que aumenta de forma notable con el aumento del grado de obesidad y con la edad. Sin embargo, cabe destacar que los resultados beneficiosos de la detección precoz y tratamientos eficaces han conseguido una mayor supervivencia que se refleja en un aumento de la prevalencia en los estudios epidemiológicos. Un diagnóstico temprano es vital, porque si se retrasa durante un tiempo prolongado, pueden surgir complicaciones como deficiencias visuales, úlceras en las extremidades, enfermedades cardíacas o accidentes cerebrovasculares. En México la principal causa de ceguera adquirida es por la retinopatía diabética, la DMT2 también es la principal causa de amputación y discapacidad. Igualmente, la etiología más común de individuos con enfermedad renal crónica es la nefropatía diabética (Simon Barquera et al. 2013).

3.1.2 Tratamiento y complicaciones.

La piedra angular del tratamiento de la DMT2 es la promoción de un estilo de vida como: la reducción de peso, una dieta individualizada con alto contenido en fibra (20-30 gr por día), ejercicio regular, y dejar de fumar (American Diabetes Association 2021b). Si los intentos de cambiar el estilo de vida no son suficientes para controlar niveles de glucosa en sangre, se suele iniciar la medicación oral, siendo la metformina el medicamento de primera línea (Larry Jameson et al. 2018). Si el tratamiento de primera línea no es suficiente o tiene contraindicaciones, se dispone ahora de un abanico de opciones terapéuticas de segunda línea como monoterapia o medicamentos combinados dependiendo de las necesidades del individuo (por ejemplo, sulfonilureas, inhibidores de la alfa glucosidasa, tiazolidinedionas, inhibidores de la dipeptidil peptidasa 4 (DPP-4), agonistas del péptido similar al glucagón 1 (GLP-1, del inglés *glucagon-like peptide-1*) e inhibidores del SGLT2). En ciertos pacientes, puede ser necesaria la administración por vía subcutánea de insulina para controlar la hiperglucemia, si los medicamentos orales no consiguen controlar los niveles óptimos de glucosa (American Diabetes Association 2021a).

Dentro las intervenciones no farmacológicas que también han sido demostradas su efectividad, resaltan la necesidad de educación y empoderamiento del paciente sobre su enfermedad. El manejo individualizado del paciente se debe adaptar a las preferencias personales del individuo, comorbilidades, polifarmacia, discapacidades (déficit visual), y acceso a recursos o atención (Chatterjee et al. 2018) Se piensa que en un futuro la microbiota intestinal tendrá un peso significante en la recomendación de dietas personalizadas con el objetivo de coadyuvar un adecuado control glucémico (Leshem, Segal, and Elinav 2020).

En los individuos con DMT2 se recomienda ampliamente un esquema de vacunación completo que incluye la vacuna de influenza estacional, hepatitis B, vacuna neumocócica y COVID-19. Estas recomendaciones se basan en la desregulación del sistema inmune que tienen los pacientes con DM que predisponen a un aumento en la severidad y susceptibilidad a contra estas infecciones que infieren protección por las vacunas (Hodgson et al. 2015). El sueño también es importante en los pacientes con diabetes, se ha encontrado que individuos con un duración del sueño menor de seis horas o más de 10 horas tienen mayor riesgo de tener obesidad, por lo que se debe evaluar constantemente esta variable, junto con la salud psicosocial del individuo (Barone and Menna-Barreto 2011).

La atención a la diabetes debe ser brindada por ambiente multidisciplinario que incluya nutriólogo, enfermeros, médicos de primer contacto, médicos especialistas, trabajadores sociales, psicólogos y rehabilitadores. Además, se debe incluir en la atención médica un seguimiento regular (al menos anualmente) y junto el tratamiento de las comorbilidades (como obesidad, HAS, dislipidemia, síndrome de apnea obstructiva del sueño, entre otras). Lo anterior tiene como objetivo evitar la aparición de enfermedades cardiovasculares ateroscleróticas, que son la principal causa de mortalidad en pacientes con DMT2. El manejo de estas comorbilidades, se puede requerir el uso de los siguientes fármacos, dependiendo de las necesidades del individuo: estatinas, inhibidores de la enzima convertidora de angiotensina o bloqueadores de los receptores de la angiotensina y aspirina en dosis bajas (American Diabetes Association 2021b).

Un inadecuado control de la hiperglucemia a largo plazo puede causar daños en muchos órganos del cuerpo, lo que provoca complicaciones microvasculares (por ejemplo, retinopatía diabética, nefropatía diabética, neuropatía diabética) y las complicaciones macrovasculares (por ejemplo, accidente cerebrovascular (ACV), enfermedad arterial

periférica (EAP)) que son incapacitantes y potencialmente mortales (Mohan and Pradeepa 2017).

3.1.3 Situación actual en México.

México se encuentra en los primeros 10 países con mayor número de adultos (20-79 años) con diabetes, con un estimado de 14.1 millones en el 2021, y se proyecta para el 2045 con 21.2 millones de personas con diabetes. Además es el sexto país con mayor número de personas adultas con diabetes no diagnosticadas, con un número estimado de alrededor 6.7 millones de personas (Tönnies et al. 2021)

La prevalencia en México de casos conocidos es de 10.3% de acuerdo con la Encuesta Nacional de Salud y Nutrición (ENSANUT) 2018, sin embargo, según de Federación Internacional de Diabetes (IDF, del inglés *International Diabetes Federation*) se estima una prevalencia alrededor de 16.9% (Intervalo de confianza (IC) del 95% de 14.2-22.1) (IDF, 10th). Además, en México alrededor del 63.4% de la población tiene sobrepeso, y 27.6% de la población es obesa, siendo el segundo lugar a nivel mundial en el reporte del año 2016 por la Organización Mundial de la Salud (OMS) (“Country Profiles” 2016).

La mayoría de los estudios prospectivos sobre la mortalidad en pacientes diabéticos se enfocan en países de alto ingreso económico. No obstante, se reporta que países en vías de desarrollo con un ingreso económico bajo o medio, como México, reportan hasta el doble la tasa de mortalidad por cualquier causa al compararlo con países de alto ingreso económico (Alegre-Díaz et al. 2016). Esto es de importancia porque se calcula que 3 de cada 4 adultos con DM viven en países con bajo y medio ingreso económico (“Country Profiles” 2016).

En un estudio prospectivo realizado en la Ciudad de México, donde se reclutaron aproximadamente 50,000 hombres y 100,000 mujeres de 35 años o más durante el período de 1998-2004, encontraron una prevalencia de DMT2 del 3% en personas con 35-39 años, y mayor al 20% en las personas de 60 años durante el estudio. En este estudio los autores reportaron que los individuos con DMT2 tenían un control glucémico inadecuado al momento de captación con valores altos de Hb glicosilada (media del 9.0%, Desviación estándar (DE) +-2.4), a pesar de que más del 90% de los pacientes se encontraba con manejo anti-diabético. En el seguimiento a los 12 años, el diagnóstico previo de DMT2 se asoció con una tasa de mortalidad por cualquier causa de 5.4 (IC del 95%, 5.0-6.0) a los 35-59 años, de 3.1 (IC del 95%, 2.9-3.3%) de 60-74 años, y de 1.9 (IC del 95%, 1.8-2.1) a los 75-85 años, que es casi el doble al compararlo con países con alto ingreso económico. Este exceso de mortalidad en individuos de 35-74 años, fue principalmente por enfermedades renales, enfermedades cardíacas, infecciones y crisis diabéticas agudas (Alegre-Díaz et al. 2016). Los resultados comparados con países de alto ingreso económico, denota un pronóstico mucho peor en los países como México, que los intentos mundiales por atender este problema empiezan a enfocar su atención en estas regiones (IDF, 10th).

A pesar de los esfuerzos realizados para modificar la historia natural de la enfermedad, la DM es uno de los principales motivos de muerte a nivel nacional causando 151 mil 214 defunciones (13.9%) según el censo 2020 del INEGI, ocupando el tercer puesto después de enfermedades del corazón y COVID-19 (Mejía et al. 2021). Por lo que es fundamental implementar políticas hacia la prevención primordial y primaria que reduzcan las tasas de obesidad, promover la actividad física en la comunidad, un mayor acceso a centros clínicos multidisciplinarios, mayor acceso a alimentos no procesados nutritivos, y se impulse la participación de los miembros de la familia y la comunidad (Bhattacharya and Roy 2016).

En México se han realizado 4 encuestas de salud poblacional en las últimas dos décadas: 2006, 2012, 2016 y 2018-2019 (Shamah-Levy et al. 2019). Estas encuestas han sido una guía fundamental para entender la situación actual de México e implementar recomendaciones para enfrentar los retos actuales del país. A pesar de eso, hay aspectos en la epidemiología de los mexicanos con DMT2 que necesitan ser explorados como la incidencia de DM en grupos especiales (incluyendo niños y adolescentes, mujeres embarazadas, grupos indígenas, DMT1, entre otros), el progreso de pacientes con pre-DMT2 a DMT2 en México, el perfil de microbiota intestinal en pacientes mexicanos a nivel nacional, entre otros.

Para resolver la epidemia de obesidad y DM en México se han realizado múltiples esfuerzos por el gobierno como: la implementación de un sistema de etiquetado frontal de la comida para identificar fácilmente productos sanos y no sanos por la población (White and Barquera 2020), campañas masivas de comunicación sobre la enfermedad y sus riesgos, regulación de la distribución de comida en las escuelas, impuestos en bebidas azucaradas y comidas procesadas, entre otros (Colchero, Molina, and Guerrero-López 2017).

Con estas acciones ha surgido la necesidad de evaluar el alcance e implementación de estas acciones. Por ejemplo, Yazmín Hugues et. reclutaron 119 centros escolares donde se reportó una participación del 15.1% de los centros (IC 95%, 9.2-22.8) en donde se implementó la guía nacional para distribuir comidas y bebidas procesadas en las escuelas. Este sondeo permitió concluir que en estas escuelas los alimentos disponibles en sus comedores escolares solo el 1% se apegaba a la guía nacional y 8.9% de los platillos elaborados se realizaban con alimentos no recomendados por la guía nacional (Hugues et al. 2021) Hay varios retos que impactan el alcance de estas medidas para frenar la obesidad y diabetes, como la interferencia de la industria alimenticia en la políticas públicas, aumento de la comida procesada (con alto contenido en grasas, azúcares y sal) de bajo costo, mayor

comercialización y disponibilidad de la comida procesada, falta de educación sobre la nutrición en la comunidad, entre otros (S. Barquera, Campos, and Rivera 2013).

Entre las estrategias de prevención primaria más efectivas, la lactancia materna se propone como una de las mejores intervenciones para prevenir la obesidad y diabetes en el adulto. La OMS recomienda una lactancia materna exclusiva durante los primeros 6 meses y posteriormente continuar hasta 2 años o más la lactancia complementaria. En México se incrementó la prevalencia de lactancia exclusiva del 13% a 20.7% del 2009 al 2018, a pesar de esto, es necesario seguir implementando estrategias y apoyo para una adecuada práctica en grupos de riesgo como mujeres solteras, mujeres trabajadoras, y/o mujeres indígenas. Por lo que se debe fomentar áreas de trabajo y zonas en la comunidad que promuevan la lactancia materna (Unar-Munguía et al. 2021)

3.1.4 Fisiopatogenia de la DMT2.

La DMT2 es una enfermedad compleja y heterogénea que se caracteriza por niveles altos de glucosa en sangre, resistencia a la insulina y una alteración en la secreción de la insulina. La patogenia de la enfermedad es multifactorial e involucra la interacción de factores genéticos y ambientales que tienen distintos grados de contribuciones en la patogenia de la enfermedad (Stumvoll, Goldstein, and van Haeften 2005). Los factores ambientales desempeñan un papel fundamental para el desarrollo de la resistencia a la insulina, los principales factores descritos son obesidad y el estilo de vida (baja actividad física, disrupción en los patrones de sueño, dieta, y microbiota intestinal). De forma general, las alteraciones genéticas contribuyen principalmente en la alteración de la secreción de insulina y también a la programación durante la etapa fetal de las células β-pancreáticas. Ambos factores ambientales y genéticos conducen a un déficit de la acción o secreción de la

insulina que sobrepasa las necesidades del cuerpo para mantener niveles normales en sangre (Melmed et al. 2015).

La progresión de la DMT2 se asocia a otras anomalías como la hipertensión y dislipidemia debido a un aumento de los ácidos grasos libres circulantes, liberación de citocinas pro-inflamatorias y factores oxidativos que en conjunto predisponen a un mayor riesgo de sufrir enfermedades cardiovasculares (como ECV, enfermedad arterial coronaria, y EAP) (Grundy 2006).

En los siguientes párrafos abordaremos a profundidad los factores cardinales en la patogénesis de la enfermedad.

3.1.4.1 Resistencia a la insulina periférica.

Mediante varios estudios prospectivos se ha encontrado a los factores ambientales como determinantes en la resistencia a la insulina, y con menor contribución factores genéticos. La resistencia a la insulina se puede detectar años antes de la instauración de la enfermedad, este fenómeno se puede observar en los individuos con pre-DMT2. Por lo cual, conocer los mecanismos que contribuyen a la resistencia a la insulina es fundamental para evitar el inicio y progresión de la enfermedad. La resistencia a la insulina puede ser causada por distintos mecanismos celulares y moleculares que incluye el aumento de ácidos grasos libres en sangre, aumento de la expresión de citocinas pro-inflamatorias y adipocinas (mediadores liberados por los adipocitos), también se involucra la activación de vías de estrés celular (Muoio and Newgard 2008).

La resistencia a la insulina se asocia con los niveles elevados de ácidos grasos libres en circulación. Ya que este aumento ocasiona una alteración en la fosforilación del sustrato del receptor de la insulina (IRS)-1. IRS-1 es una proteína adaptadora fundamental para inducir la

expresión de los canales de glucosa (GLUT-4). Además, este exceso de lípidos en pacientes con DMT2 causa su acumulación en las mitocondrias y alterando la señalización de la insulina, principalmente en el músculo esquelético (Koenig et al. 2019).

La inflamación sistémica de bajo grado es un componente en la patogénesis de la DMT2 y del desarrollo de complicaciones vasculares a largo plazo. Un componente que influye en la inflamación sistémica de los pacientes con DMT2 es la disbiosis intestinal. Este fenómeno se asocia a que los pacientes con DMT2 tienen un aumento de la permeabilidad intestinal, que permite el paso de lipopolisacáridos (LPS) de origen bacteriano a la circulación sistémica (Scheithauer et al. 2020). Los LPS son reconocidos por los receptores del sistema inmune innato y su reconocimiento induce la expresión de citocinas pro-inflamatorias como factor de necrosis tumoral-α (TNF-α), interleucina-1 (IL-1), IL-6, entre otros. El TNF-α causa resistencia a la insulina al reducir la fosforilación de la tirosina en el receptor de la insulina y el IRS-1. Además, el TNF-α induce fosforilación en la serina del IRS-1, en conjunto estas modificaciones causan una disminución de la transducción de señales río abajo de la insulina (León-Pedroza et al. 2015).

3.1.4.2 Secreción de la insulina

La secreción defectuosa de insulina puede ser causada por diferentes factores de riesgo. Entre los más importantes destacan los factores genéticos (HNF1A, TCF7L2, MTNR1B) y la desregulación de la programación intrauterina de las células β del páncreas (Flannick et al. 2019). En relación, los propios componentes de la enfermedad pueden contribuir a la alteración en la secreción de la insulina. La hiperglucemia puede deteriorar la función de las células beta pancreáticas mediante un mecanismo conocido como glucotoxicidad, que causa un daño directo a las células por los niveles elevados de glucosa.

Un mecanismo importante es el aumento compensatorio en la producción de pro-insulina y pro-amilina ocasionado por la resistencia a la insulina (Xie et al. 2022). La acumulación de estas proteínas en los islotes del páncreas, como pro-amilina, ocasiona una alteración en la producción de insulina. La acumulación ocurre debido a que las enzimas proteolíticas no son capaces de responder a la sobredemanda de glucosa o hiperglucemia. Los depósitos de amilina (compuestos por oligómeros de amilina y fibrillas de amiloide) también contribuyen al daño del páncreas en los pacientes con DMT2. Este fenómeno, se ha observado en pacientes con demencia tipo Alzheimer, pero en el cerebro y son conocidas como placas de beta amiloides neuronales (Xie et al. 2022).

Estos factores ambientales, genéticos y propios de la enfermedad (incluyendo la hiperglucemia y la resistencia a la insulina) ocasiona arcos de retroalimentación en la patogenia de la enfermedad. En conjunto perpetúan el estado metabólico alterado de los pacientes con DMT2 (Koenig et al. 2019).

3.1.4.3 Susceptibilidad genética y epigenética

La DMT2 es una enfermedad poligénica que interactúa con factores en el ambiente y factores epigenéticos. Se han descrito más de 500 variantes genéticas asociadas a la DMT2, la mayoría se encuentra relacionada con la función de las células beta pancreáticas y en la acción de la insulina (Vujkovic et al. 2020). El 90% de los individuos con DMT2 tienen al menos un familiar con diabetes, y en caso de tener un familiar de primer grado con DMT2 se tiene hasta el doble de riesgo de desarrollar la enfermedad (Riesgo relativo =2.24, valor $P<0.0001$) (Weires et al. 2007). Cabe destacar que la presencia de múltiples polimorfismos en un individuo eleva la posibilidad de desarrollar DMT2, sin embargo, variantes monogénicas tienen un efecto muy pequeño en el riesgo. Por lo que la combinación de las variantes genéticas junto con factores ambientales es necesaria para acercarse a una predicción correcta de la DMT2 (SIGMA Type 2 Diabetes Consortium et al. 2014).

A pesar de que varios estudios han descrito un elevado riesgo de tener la enfermedad al tener algún familiar con DMT2. Este riesgo podría deberse a mecanismos independientes de la genética como el entorno donde viven y acceso a recursos de la familia o comunidad. También se debe a las marcas epigenéticas nocivas que pueden ser consecuencia de la sobre-alimentación y diferentes "inadecuados" hábitos en el estilo de vida en los individuos (Rosen et al. 2018).

Principalmente, se ha hipotetizado que la malnutrición (defecto o exceso) durante la etapa fetal y temprana de la vida, ocasiona un cambio en la programación metabólica que se asocia con un riesgo incrementado de desarrollar síndrome metabólico en la vida adulta (que incluye resistencia a la insulina, DMT2, dislipidemia, obesidad, e hipertensión) (BIRTH-GENE (BIG) Study Working Group et al. 2019). Se ha demostrado en modelos de ratón que la sobre-nutrición durante la etapa fetal, aumenta el riesgo de obesidad en los hijos y que se extiende el riego a siguientes generaciones (Seki et al. 2012).

3.2 Microbiota intestinal

La microbiota humana se refiere al conjunto de microorganismos (bacterias, archaea, hongos, y virus, que cuentan aproximadamente 90 trillones en total) que residen en el cuerpo humano. En cambio, el microbioma humano abarca a los genes y productos genéticos (ARN, proteínas y metabolitos) producido por las comunidades microbianas residentes. La composición del microbioma varía dependiendo de la región anatómica del cuerpo humano (Marchesi and Ravel 2015). Entre las regiones más diversas y abundantes respecto a la composición del microbioma humano se encuentra la que reside en el lumen o luz intestinal. La microbiota intestinal codifica más de 3 millones de genes, y se modifica de forma con la dieta, enfermedad, y estilo de vida del individuo. El cambio dinámico que

experimenta la microbiota permite a la persona adaptarse a las necesidades cambiantes del entorno que habita (Aagaard, Luna, and Versalovic, n.d.).

La microbiota intestinal es fundamental para el ser humano desde los primeros años de vida. Participa en la digestión, degradación de toxinas, regulación endocrina, modulación y maduración de la respuesta inmune, entre muchas otras (Flint et al. 2012). Por lo que las funciones de la microbiota intestinal no se limitan únicamente al lugar donde residen (lumen o luz intestinal), sino que también pueden afectar de forma sistémica al individuo. Esto es de importancia, porque la alteración en la estructura de la microbiota conocida como disbiosis puede asociarse con enfermedades, tanto locales como sistémicas. Entre los padecimientos asociados a la disbiosis intestinal se encuentran trastornos intestinales (e.g. síndrome intestino irritable), enfermedades autoinmunes (e.g. DMT1, EII, enfermedad celíaca), cáncer y diversas enfermedades metabólicas (e.g. obesidad y DMT2) (Marcos-Zambrano et al. 2021).

Como mencionamos anteriormente, la microbiota intestinal tiene la capacidad de actuar de forma local y sistémica. Esta función lo logra mediante la generación de metabolitos que actúan en el sitio de producción (llamada vía paracrina) y en diferentes tejidos periféricos que se transportan a la circulación sistémica (llamada vía endocrina) y modifican la fisiología del individuo. Entre los metabolitos más importantes producidos por la microbiota intestinal se encuentra la generación de ácidos grasos de cadena corta (AGCC), biotransformación de ácidos biliares (AB) secundarios, la formación de trimetilamina (TMA), entre otros (Sharma and Tripathi 2019). Cabe resaltar, que se ha encontrado que la microbiota intestinal tiene la capacidad de generar neurotransmisores como GABA o glutamato (Strandwitz et al. 2019). Estos efectos pleiotrópicos por parte de la microbiota intestinal son un ejemplo de su capacidad de modular de forma general la salud del cuerpo humano, mostrando la

capacidad de la microbiota intestinal para modular el metabolismo celular, regulación hormonal y permitir una adecuada función del sistema inmune.

Para que haya una adecuada respuesta inmune es esencial que se mantengan saludables las barreras del cuerpo humano, entre estas se encuentran las mucosas que recubren las cavidades del cuerpo humano. La mucosa en el sistema gastrointestinal se considera la más extensa y esencial para que el sistema inmune responda contra infecciones. La mucosa en el sistema gastrointestinal se integran varios componentes, entre estos se encuentra un tejido especializado rico en células del sistema inmune que se conoce como tejido linfoide asociado a mucosa o MALT (del inglés, *Mucosa-associated lymphoid tissue*). Este componente es necesario para responder contra infecciones potenciales que ingresan al sistema gastrointestinal y mantener una adecuada homeostasis entre la microbiota-huésped. La región más importante del MALT se encuentra en la lámina propia de la parte terminal del íleon (porción distal del intestino delgado) intestinal y se conoce como Placas de Peyer. Esta región se encuentra formada por una zona rica en linfocitos B conocida como folículos linfoides y alrededor de ella cuenta con diversas células del sistema inmune como linfocitos T CD4+ o cooperadores (principalmente TH17 y FOXP3+), células linfoides innatas, macrófagos residentes, células dendríticas, entre otros (Hino et al. 2020).

La microbiota es esencial para la maduración o también llamada “entrenamiento” del sistema inmune en los primeros años de vida y el mantenimiento de la barrera intestinal en la vida adulta (figura 1). Una disbiosis intestinal que influye en la alteración de estas funciones se encuentra asociada a múltiples patologías del individuo como obesidad, diabetes, asma bronquial, alergia, cáncer, entre otras (Al Nabhani and Eberl 2020).

En los primeros años de vida es necesario la participación de la microbiota intestinal para el desarrollo y maduración de las placas de Peyer. Esto fue demostrado cuando en un ratón

deficiente del receptor NOD-1 que tiene la capacidad de reconocer peptidoglicano de la microbiota, no tuvo la capacidad de formar adecuadamente folículos linfoides de las placas de Peyer. Este estudio concluyó que la interacción NOD-1/peptidoglucano es necesaria para producir las quimiocinas CCL19 y CCL20. Ambas son necesarias para la migración de los linfocitos a la mucosa intestinal (Bouskra et al. 2008). También se ha demostrado que los AGCC de la microbiota intestinal de la madre embarazada inducen la proliferación y supervivencia de los linfocitos T reguladores FOXP3+ en MALT. Los linfocitos T reguladores mantienen la tolerancia periférica contra antígenos provenientes de la dieta, y es indicativo de una mucosa saludable. Los pacientes con disbiosis intestinal tienen mayor incidencia y riesgo de enfermedades alérgicas como asma bronquial o alergia alimentaria (Al Nabhani et al. 2019)

Para que haya una adecuada función intestinal es necesario que su microbiota se mantenga contenida por el sistema inmune. Esto es logrado al mantener una adecuada función de la barrera intestinal y sus componentes. El moco producido por las células caliciformes del intestino depende de la expresión de mucinas secretadas y mucinas transmembranales (como MUC1, MUC3, MUC17). La regulación de la expresión de estas mucinas está sujeta a la producción de citocinas generadas (IL-13, IL-22, IL-17) por parte del sistema inmune. Otro componente fundamental es la integridad del epitelio, los coloncitos se mantienen impermeables gracias a la expresión de sus uniones estrechas. La microbiota intestinal mantiene esta estructura mediante los distintos mediadores que produce como los AGCC y al estimular la producción de la IL-17 por el sistema inmune (Al Nabhani and Eberl 2020).

Finalmente, la producción de Inmunoglobulina A (IgA) dimérica también es fundamental para mantener una barrera intestinal sana. Esta inmunoglobulina es necesaria para neutralizar toxinas y microorganismos potencialmente patógenos, y su expresión es regulada en parte por la microbiota intestinal. La producción de IgA depende de los linfocitos B que se

encuentran en las placas de Peyer, ya que al reconocer un antígeno se activarán y diferenciarán en células plasmáticas productoras de IgA dimérica (Seki et al. 2012). Las bacterias tienen diferentes componentes como LPS y peptidoglicano en su estrucutra que al ser reconocido por las células dendríticas provocan un aumento la producción de factores de supervivencia (incluyendo el ligando inductor de proliferación (APRIL, por sus siglas en inglés), factor de activación célula B (BAFF, por sus siglas en inglés) que sirven para mantener a los linfocitos B maduros productores de IgA (Al Nabhani and Eberl 2020).

La composición de la microbiota cambia en función de varios factores como la edad, la dieta, la enfermedad, el medio ambiente, y el estilo de vida. Estos cambios influyen en la función inmunitaria local de mucosas y también a nivel sistémico. Consideramos que conocer los mecanismos anteriores por los cuales la microbiota intestinal influye en la respuesta inmune es fundamental para entender cómo influye la DMT2 a la disbiosis intestinal, y a la inversa.

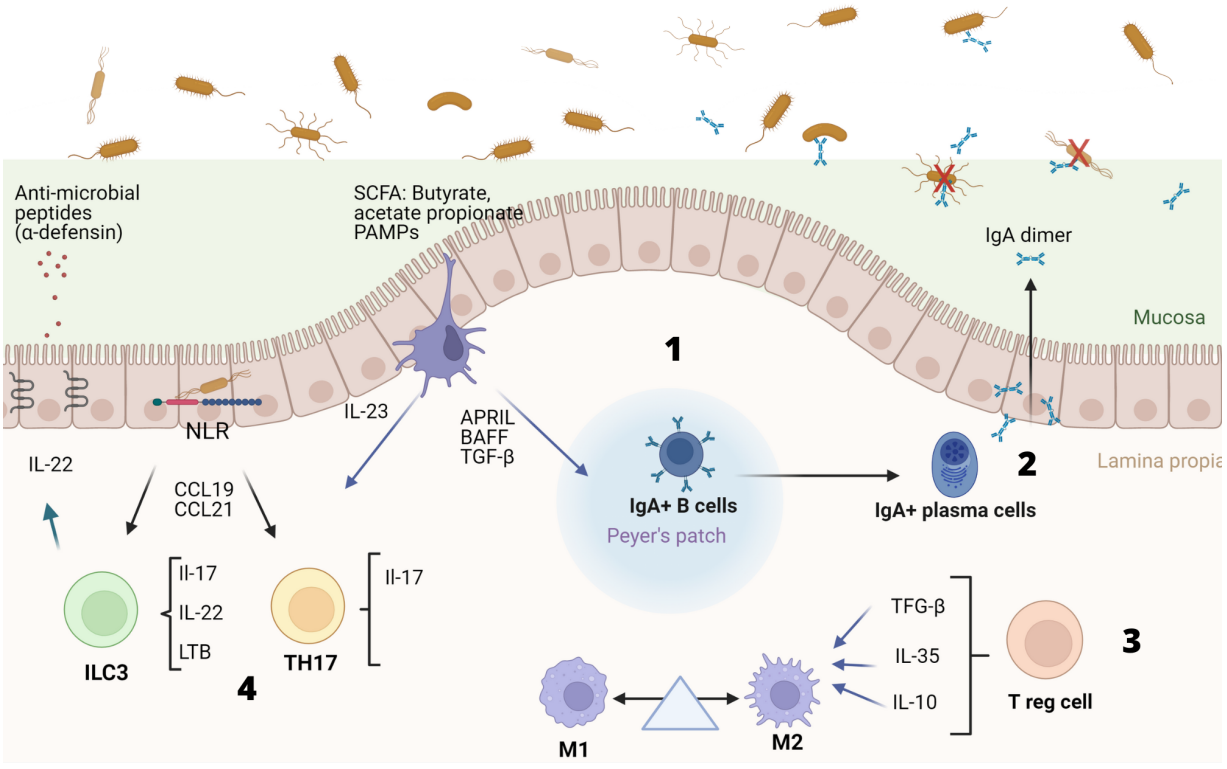

Figura 1. Esquema de modulación del sistema inmune por parte de la microbiota intestinal.

1) Mantenimiento del tejido linfoide asociado en mucosas (MALT, por sus siglas en inglés) a través de la estimulación del sistema inmune por parte de la microbiota intestinal para liberar factores de supervivencia (BAAF y APRIL) para los linfocitos B maduros. 2) Mantenimiento de la integridad epitelial (expresión de ocludina y claudina) y producción de moco (expresión de MUC1) a través de la estimulación por mediadores como butirato y citocinas del perfil TH17 (IL-17, IL-22). 3) Estimulación de la producción y secreción de IgA diméricas al lumen intestinal, que mantienen la homeostasis intestinal. 4) Quimioatracción de los Linfocitos TReg (FoxP3 +) a la mucosa para mantener la tolerancia oral contra antígenos no dañinos provenientes de la dieta. APRIL: A proliferation-inducing ligand; BAAF: B cell–activating factor; CCL: C-C Motif Chemokine Ligand; Ig: immunoglobulin; IL: interleukin; NLR: Nod-like receptor, PAMPS: pathogen-associated molecular pattern molecules ; SCFA: Short chain fatty acids; TGF: transforming growth factor.

3.3 Microbiota intestinal en pacientes con Diabetes Mellitus tipo 2

Desde hace un lustro, se han descrito varios estudios que proporcionan una evidencia clara sobre una asociación directa entre la microbiota intestinal y la patogenia de la DMT2. Motivados por estos estudios, se han tratado de dilucidar los mecanismos de esta contribución en el progreso de la enfermedad y entender como el estado diabético influye en la composición de la microbiota intestinal. Algunas consecuencias inmediatas de esta disbiosis intestinal en pacientes con DMT2 se ha descrito un aumento del número de patógenos oportunistas y una disminución de las especies productoras de ácidos grasos de cadena corta (AGCC) (Seki et al. 2012; Sanna et al. 2019).

De la misma forma se ha observado una asociación entre cambios en la composición de la microbiota intestinal con la resistencia a la insulina (Lee, Sears, and Maruthur 2020). Los principales cambios que ocurren en los pacientes con DMT2 se relacionan con vías inmuno-inflamatorias que modifican la producción de AGCC y la biotransformación de

ácidos biliares (AB) secundarios por parte de la microbiota. También se ha identificado que los pacientes con DMT2 presentan cambios en el metabolismo de aminoácidos de cadena ramificada y las vías de reducción de sulfato por parte de la microbiota intestinal (Sharma and Tripathi 2019).

La microbiota se encarga de producir AGCC a través de la fermentación de la fibra dietética, como butirato, acetato y propionato (Martin-Gallausiaux et al. 2021). Estos AGCC sirven como fuente de energía para los colonocitos en el hospedero, los cuales permiten mantener un fenotipo normal propiciando la integridad epitelial y evitando el paso de toxinas y/o patógenos. Además, los AGCC también tienen funciones moduladoras en el metabolismo celular y sistema inmunológico (Yao et al. 2022). Entre estos mediadores, destaca butirato que mejora la sensibilidad de la insulina al unirse a distintos receptores transmembranales en los enterocitos como el receptor acoplado a la proteína (GPR, del inglés *G protein-coupled receptor*) tipo 41, y el GPR43. Esta unión ligando-receptor induce la liberación de las incretinas como el GLP-1, por sus siglas en inglés) y el péptido YY. Las incretinas son hormonas que estimulan la liberación de insulina, inhiben la secreción de glucagón, inducen saciedad y mejoran la sensibilidad de insulina, figura 2A. En diferentes estudios de pacientes con DMT2 se ha demostrado una disminución de especies productoras de AGCC como *Lactobacilli, Roseburia, Faecalibacterium prausnitzii, y Bifidobacteria (den Besten et al. 2015)* y por lo tanto se hipotetiza que podría estar relacionado con los altos niveles de glucosa y alteraciones en la señalización de la insulina que caracteriza al paciente diabético.

Por otro lado, se ha reportado una disminución de la concentración de AB secundarios en pacientes con DMT2. Los AB primarios son sintetizados en el hígado, y posteriormente son biotransformados por la microbiota intestinal a AB secundarios entre los que se encuentra el ácido hiodeoxicólico y el ácido litocólico. Los AB secundarios son importantes porque ayudan a modular el metabolismo de la glucosa y el de lípidos. Las acciones de los AB

secundarios dependen de los receptores membranales, como el receptor de AB acoplado a proteína G Takeda (TGR5) y, los receptores nucleares, como el receptor X fernesoide (FXR) que se expresan en varios tejidos como el íleon, colon, y tejido adiposo. Específicamente el receptor TGR5 se encuentra expresado en células L del epitelio intestinal, promueve la producción del GLP-1 que estimula la producción de glucosa dependiente de insulina. (Pols et al. 2011) Bajo este contexto, la disminución en la producción de los AB secundarios debido a la disbiosis intestinal podría influir en la alteración en la secreción de insulina en los pacientes con DMT2 (figura 2B).

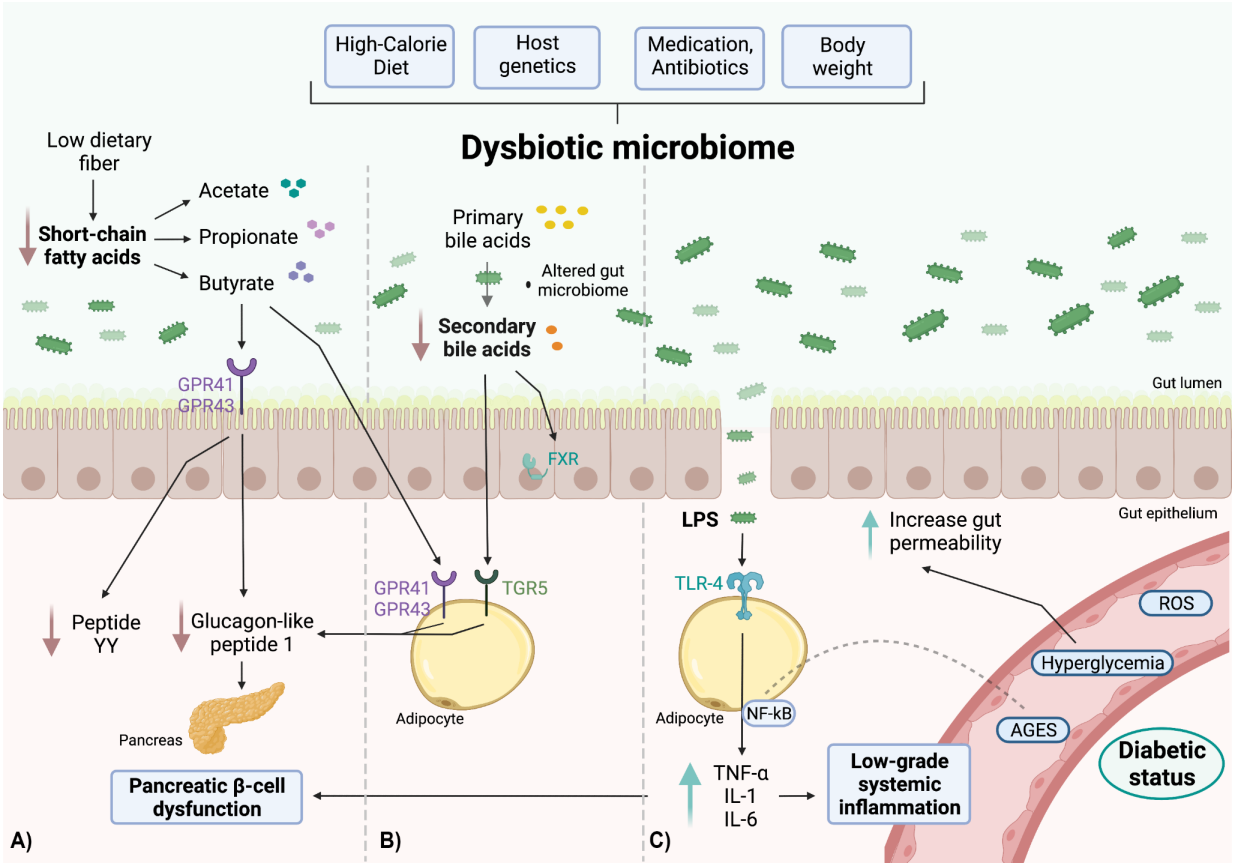

Figura 2. La disbiosis intestinal contribuye a la progresión de la DMT2. A) Hay una disminución de la producción de AGCC (como el butirato, acetato y propionato), estos productos derivados de la fibra dietética se unen a los receptores GPR1 y GPR4 estimulando la secreción de GLP-1 y péptido YY por parte de las células L intestinales. **B)** Con una disbiosis intestinal hay una disminución en la

biotransformación de los ácidos biliares secundarios, que generalmente estimulan la secreción de GLP-1 en la células L intestinales al activar sus receptores celulares (como FXR y TGR5) **C)** Hay un incremento de la permeabilidad intestinal debido a efecto directo glucotóxico y una disbiosis intestinal. Ocasionando un mayor probabilidad de que los lipopolisacáridos (LPS), se unen al receptor TLR4 presente en adipocitos, monocitos, células epiteliales. Esta interacción TLR4/LPS ocasiona la activación de vías de señalización que conducen a un incremento en la expresión de citocinas proinflamatorias (TNF-a, IL-6, IL-1) características del estado inflamatorio sistémico de bajo grado en pacientes con Diabetes mellitus tipo 2. (Imagen obtenida de Yoscelina Estrella et al., 2022).

La integridad de la barrera epitelial del intestino es fundamental para evitar el paso de microorganismos y toxinas hacia la circulación sistémica. En los pacientes con DMT2 se ha descrito el fenómeno de endotoxemia metabólica, que se refiere a la presencia en el torrente circulatorio de lipopolisacáridos (LPS) o también llamadas endotoxinas (figura 2C). La aparición de LPS en la sangre ocasiona la activación del sistema inmune manteniendo un estado de inflamación sistémica de bajo grado en el individuo. Se ha propuesto que la endotoxemia metabólica, se deba principalmente a un incremento de la permeabilidad intestinal (conocido también como 'fuga epitelial') que puede ser ocasionado por distintas razones (Mohammad and Thiemermann 2020). El incremento de la permeabilidad intestinal se debe principalmente a mecanismos como glucotoxicidad contra las células epiteliales, una respuesta inflamatoria local anormal en la mucosa o la disbiosis intestinal. La presencia de LPS a nivel sistémico permite que pueda ser reconocido por células inmunes, tales como células dendríticas, y no inmunes como los adipocitos ya que tienen receptores del sistema inmune innato (PRRs, del inglés) como el receptor tipo Toll 4 (TLR4, del inglés *Toll-like receptor type 4*). El reconocimiento de LPS por el receptor TLR4 ocasiona el reclutamiento de la proteína adaptadora MyD88, que a su vez concluye con la activación del factor nuclear κB (NF-κB, del inglés *Nuclear factor κB*) (Mohammad and Thiemermann 2020). La activación de NK-*κ*B ocasiona la síntesis de citocinas pro-inflamatorias como IL-1, IL-2, TNF-α, adipocinas, y

la proteína quimiotáctica de monocitos tipo 1 (MCP-1), estas últimas inician el reclutamiento y activación de las células del sistema inmune para iniciar el proceso inflamatorio. Como mencionamos anteriormente, las citocinas proinflamatorias alteran directamente en la señalización de la insulina y su presencia en sangre sirven como predictores de resistencia a la insulina, incidencia de DMT2 y enfermedades cardiovasculares (Tsalamandris et al. 2019).

La respuesta inflamatoria es importante para el inicio y formación de la placa aterosclerótica en los vasos sanguíneos, que es clave para el progreso de enfermedades cardiovasculares (incluyendo infarto agudo al miocardio, ECV) que ponen en peligro la vida del paciente con DMT2. Además, la liberación de las citocinas proinflamatorias se asocian a un déficit en la señalización de la insulina ya que intervienen de forma directa en la fosforilación del IRS-1 y el receptor de la insulina, ocasionando resistencia a la insulina (Carvalho and Saad 2013).

3.4 Algoritmos de aprendizaje de máquina para estudios del microbioma

3.4.1 Generalidades de los algoritmos de aprendizaje de máquina.

Aprendizaje de máquina o ML (Machine Learning) es una rama de la inteligencia artificial y se define como el campo de estudio que le da a la computadora la habilidad de aprender sin la necesidad de ser explícitamente programado (Naresh et al. 2020). En general se puede clasificar de forma amplia en: aprendizaje de máquina supervisado y aprendizaje de máquina no supervisado (Naresh et al. 2020).

En el caso de aprendizaje no supervisado, el objetivo es desarrollar algoritmos capaces de encontrar patrones en un conjunto de datos y agrupar observaciones en variables

numéricas, un ejemplo de ellos sería los análisis de agrupación (e.g. K-means) y las técnicas reducción de dimensionalidades (e.g. PCA, UMAP) (Naresh et al. 2020).

En el caso del aprendizaje supervisado, su objetivo es aprender a clasificar grupos de datos y realizar predicciones. Se utilizan para resolver problemas de regresión y clasificación. En el problema de regresión, se trata de predecir el resultado de una variable continua (e.g. edad, pH, índice de masa corporal). En el problema de clasificación se trata de predecir el resultado de una variable categórica: dicotómica o multinomial (e.g. estado de salud, fenotipo maligno o benigno). Entre los algoritmos de ML supervisado encontramos la regresión logística, regresión lineal, naïve Bayes, redes neuronales artificiales, máquinas de vectores de soporte, random forest, XGBoost (Libbrecht and Noble 2015).

El término aprendizaje proviene de la capacidad que tiene la máquina para resolver una tarea de forma correcta utilizando la "experiencia", consiste en actualizar los "parámetros" del modelo para obtener el mejor rendimiento (figura 3, parte a). Los "parámetros", también conocido como "pesos", son valores que modifican nuestros datos de entrada para lograr un resultado y pueden ser modificados o actualizados para mejorar el rendimiento. El rendimiento de los resultados del modelo se utiliza para evaluar al modelo y saber que tan correcto se hizo la tarea. La forma de evaluar el rendimiento del modelo generalmente se realiza al comparar la predicción o resultado del modelo contra la etiqueta o valor real. El modelo ajusta o "adapta" los "pesos" para reducir la distancia entre los resultados o predicciones y los valores reales. Una vez que ya tengamos un modelo entrenado con los "pesos" o "parámetros" óptimos, se pone a prueba su capacidad para resolver alguna tarea de clasificación o regresión (figura, parte b) (Mohri, Rostamizadeh, and Talwalkar 2018).

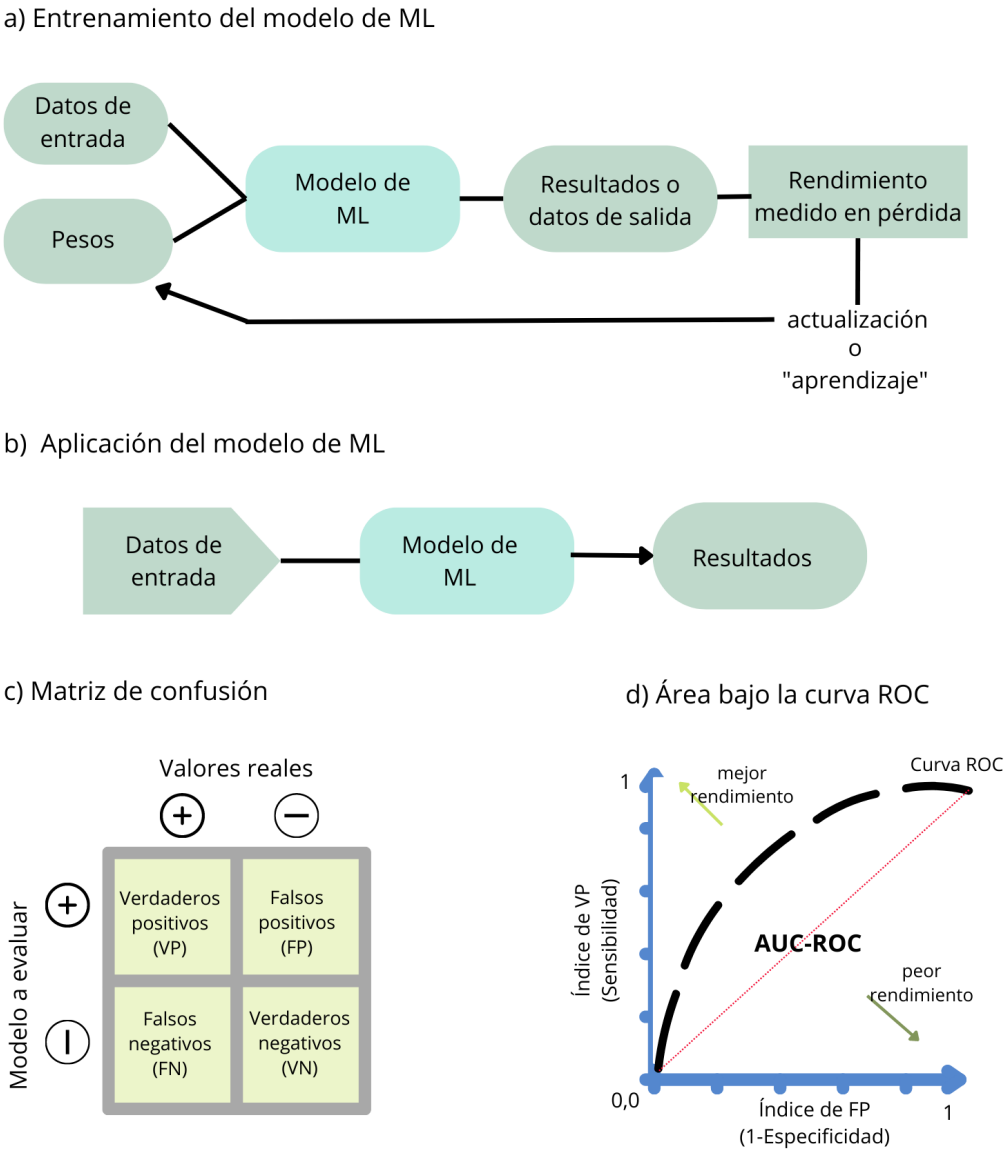

Figura 3. Generalidades de los algoritmos de aprendizaje de máquina. a) Entrenamiento del modelo de aprendizaje de máquina (ML). b) Aplicación del modelo de ML. c) Matriz de confusión. d) Área bajo la curva (AUC) ROC.

Entre los objetivos principales que tienen los métodos de ML es obtener la generalización del modelo o clasificador, que se logra al entrenar el modelo con un conjunto de datos de entrenamiento. Esta capacidad de generalización se refiere a la habilidad de clasificar

adecuadamente y poder realizar predicciones acertadas en este conjunto de datos que nunca habían visto el modelo. La generalización en ML se logra al dividir la base de datos en un subconjunto de datos de entrenamiento (generalmente entre 70-80%) y otro subconjunto de prueba (20-30%), los cuales permitirán aprender a realizar clasificaciones y a evaluar el rendimiento del modelo, respectivamente.

Una forma común de dividir la base de datos en el conjunto de entrenamiento y de prueba, es utilizar la técnica de validación cruzada (CV, *cross validation* por sus siglas en inglés) que además permite mostrar la dispersión de la precisión del modelo. La CV consiste en dividir los datos en subconjuntos de igual tamaño llamados pliegues o *k-folds*, generalmente 5 o 10 pliegues, para posteriormente usar estos pliegues para entrenar y evaluar el modelo de forma iterativa con la combinación de los pliegues. El resultado de la CV se obtiene mediante el promedio de las iteraciones realizadas, que permite mostrar la dispersión en el resultado. En caso de tener un problema de desbalance de datos (es decir cuando no se tienen proporciones equivalente en los grupos de análisis), se ha propuesto utilizar la CV estratificada *K fold*, donde se mantiene el radio entre los casos y controles de toda la base de datos al repartir los k-fold (Santos et al. 2018).

Existen ciertas distintas métricas para evaluar el rendimiento del modelo de clasificación. Entre ellas se encuentran las siguientes: precisión, sensibilidad, especificidad, exactitud, valor-F (del inglés, *F1-score*), AUC-ROC, entre otras. Cada métrica permite evaluar al modelo de distinta forma y son obtenidas utilizando los valores obtenidos en la matriz de confusión (tabla 2x2). A través de la matriz de confusión se obtienen: los verdaderos positivos (VP) cuando el modelo realiza una correcta clasificación positiva; los verdaderos negativos (VN) cuando el modelo realiza una correcta clasificación negativa; falsos positivos (FP) cuando el modelo realiza una incorrecta clasificación positiva; y los falsos negativos (FN) cuando el modelo realiza una incorrecta calificación negativa (figura 3, parte c) (Topçuoğlu et al. 2020).

La precisión es obtenida mediante la siguiente fórmula: $precisión = VP / (VP + FP)$ y explica cuántos casos clasificados como positivos son realmente positivos comparado con los valores reales. En contraste, la sensibilidad nos habla cuántos casos de la clase positiva fueron correctamente clasificados como positivos por el modelo, es obtenido mediante la siguiente fórmula: $sensibilidad = VP / (VP + FN)$. Para obtener una combinación entre estas dos métricas (precisión y sensibilidad) se utiliza el valor-F, y se logra utilizando la siguiente fórmula: $F1 = 2 \cdot \frac{(precisión \cdot sensibilidad)}{precisión + sensibilidad}$. Por otro lado, la especificidad nos dice la proporción de la clase negative que fue correctamente clasificado mediante la siguiente fórmula: $especificidad = VN / VN + FP$. En caso de querer conocer la porción de casos que el modelo ha correctamente clasificado de la clase negativa y positiva obtenemos la métrica de exactitud, usando la siguiente fórmula: $exactitud = (VP + FN) / (VP + VN + FP + FN)$. En conjunto, la utilización de estas técnicas permiten evaluar al modelo en diferentes aspectos, sin embargo, una de las métricas más utilizada que permite comparar el rendimiento entre modelos es el valor AUC-ROC (Topçuoğlu et al. 2020).

El AUC-ROC es utilizado comúnmente para evaluar la capacidad discriminativa de un clasificador binario, con rangos de valores del 0 al 1. Los valores cercanos a 1 de AUC-ROC significa que el clasificador separa de forma casi perfecta a las dos clases, y los valores por debajo de 0.5 se refiere a que el clasificador no tiene la capacidad para discriminar las clases o son resultados del azar. Para el cálculo de AUC-ROC es necesario representar de forma gráfica a la curva ROC mediante relación entre el índice de FP (probabilidad de clasificar un valor como positivo cuando es realmente negativo) en el eje de las X y la sensibilidad (probabilidad de clasificar a un valor como positivo de forma correcta) en el eje de las Y . Por lo tanto, valores altos de rendimiento en el eje de las Y (sensibilidad) se refiere cuando el modelo obtiene más VP que FN. De igual forma, cuando se obtienen valores de rendimiento

bajos en el eje de la X (índice de FP), obteniendo un menor número de FP que VN. Por lo tanto, el objetivo es tener un curva ROC con valores altos de sensibilidad acompañada de valores bajos del índice de FP que se traduce en un AUC-ROC mayor o cercana a 1 (Davis and Goadrich 2006).

Entre los problemas que sufren los algoritmos de ML se encuentra el fenómeno de sobreajuste (del inglés *overfitting)*. Se refiere cuando el modelo es muy sensible a los patrones aprendidos en los datos de entrenamiento, que ocasiona un inadecuado rendimiento en el conjunto de prueba. En contraste con su antónimo, el sub-ajuste o *underfitting* en este caso ocurre cuando el modelo es muy sencillo para modelar los datos de entrenamiento y un bajo rendimiento en el conjunto de prueba. Por lo que es fundamental encontrar un balance entre el sobre-ajuste y el sub-ajuste que permita modelar los datos de entrenamiento y obtener un rendimiento adecuado en los datos de prueba (Ying 2019).

La cantidad y calidad de datos disponible es también un factor importante para los métodos de ML. A pesar de que el crecimiento del estudio del microbioma ha sido acompañado con el crecimiento y aplicación de tecnologías de alto rendimiento que proporcionan gran información, sigue siendo un problema obtener una gran muestra de pacientes que permitan alimentar los modelos de ML. Los algoritmos simples como regresión logística necesitan menos datos en comparación con algoritmos más complejos como las redes neuronales artificiales que dependen directamente de la cantidad de datos para aprender. Por el momento no hay una cantidad específica de datos para obtener rendimientos óptimos o que te guíen hacia la elección de un modelo predictivo específico (McCoubrey et al. 2021).

3.4.1 Ejemplos de algoritmos de aprendizaje de máquina.

En el caso del estudio del microbioma, se han aplicado distintos algoritmos de ML supervisado para la predicción del fenotipo del paciente y su pronóstico. Además, estos

métodos han sido relevantes para la búsqueda e identificación de biomarcadores con cierto grado de éxito (McCoubrey et al. 2021; Borenstein and Muller 2021). Entre los distintos métodos de ML supervisados, se encuentran los siguientes: (Tabla 1)

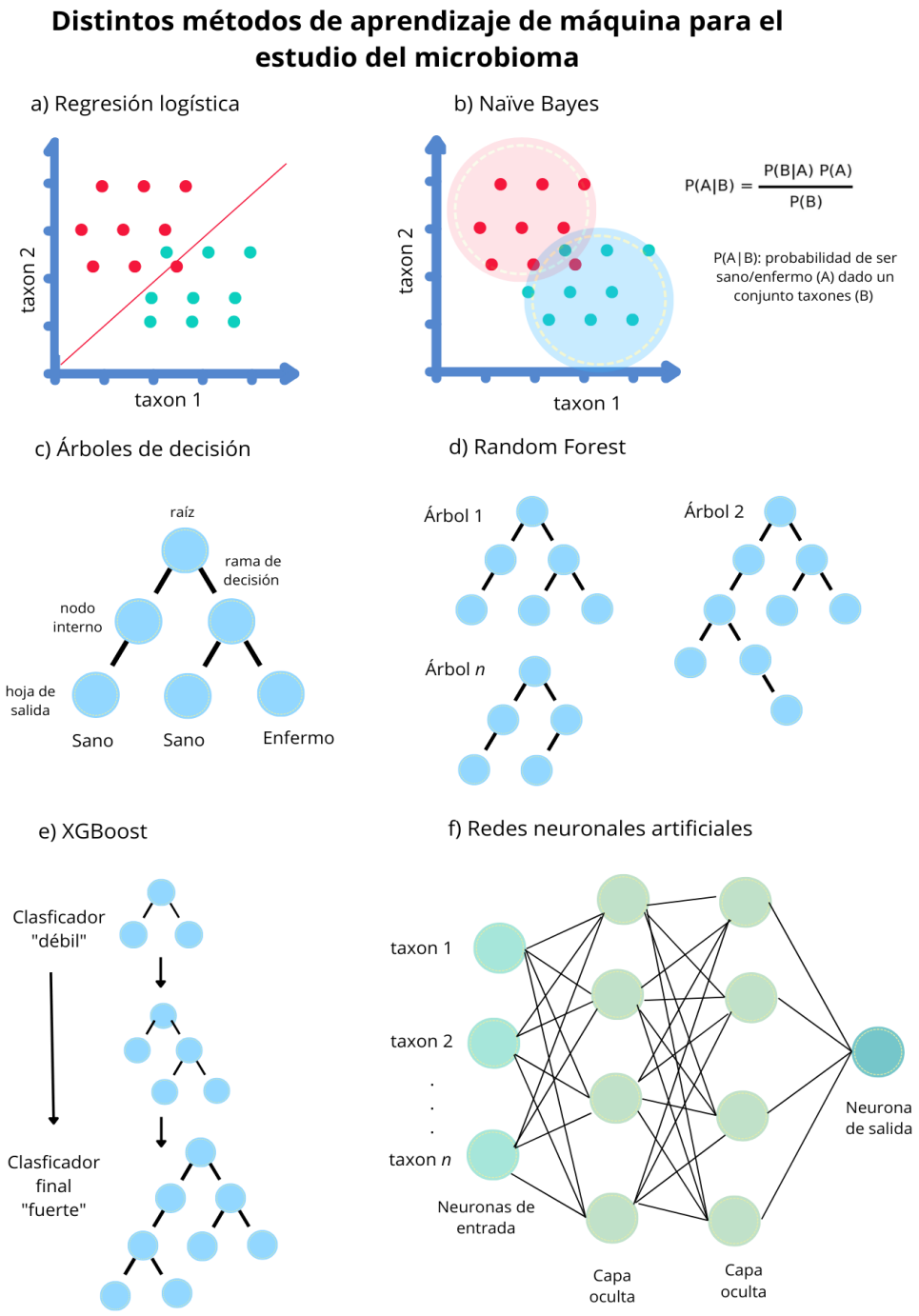

Figura 4. Ejemplos de algoritmos de aprendizaje de máquina (ML) para el estudio de la asociación microbioma-enfermedades: a) regresión logística, b) Naïve Bayes, c) árboles de decisiones, d) Random Forest, e) XGBoost, y f) redes neuronales artificiales (perceptrón multicapas)

Regresión logística binaria. Es un método linear estadístico que aprende a predecir el resultado de una variable dicotómica *Y* a partir de variables independientes (continuas o categóricas) llamadas *X*. Para lograr la clasificación se utiliza la función logística también conocida como función sigmoide para dar una probabilidad condicional dado cierto valores *X* definida como: $P(Y|X)$. Por lo tanto los rangos de probabilidad se encontrarán entre 0 y 1, si la probabilidad es mayor que 0.5 clasifica para enfermos (*Y*= 1) dado ciertos valores *X* y si la probabilidad es menor que 0.5 clasifica para sanos (*Y*= 0) dado ciertos valores *X (Bisong 2019)*.

Para lograr una adecuada clasificación el modelo de regresión logística se ajustan los valores de los parámetros tratando de reducir la función de costo. Se usa principalmente el error cuadrático medio (MSE, del inglés *mean squared error*) como función de costo, representado con la fórmula: $MSE = \frac{1}{n}\sum_{i=1}^{n}(Y - \widehat{Y})^2$. Siendo *n* el número de la muestra, *Y* la etiqueta real y $\hat{Y}$ la etiqueta predicha. Por consiguiente, la función de costo mide la distancia entre la clase actual contra la clase que predijo el modelo (Bisong 2019).

Entre los parámetros más importantes que se pueden seleccionar en el algoritmo de regresión logística se encuentra: *solver*: el algoritmo que usa para optimizar o adaptar los pesos (por defecto= "*lbfgs"*, algoritmo complejo para encontrar el valor global mínimo óptimo); *max_iter:* número máximo de iteraciones para converger por el *solver (por defecto= 100); penalty:* es una técnica de regularización para evitar el sobre ajuste (por defecto= *L2*); C: valor inverso a la fuerza de penalización (por defecto= 1) (Bisong 2019)

El algoritmo de regresión logística se ha utilizado en ciertos estudios del microbioma para identificar enfermedades. Por ejemplo, Beck and Forest, 2015 encontraron que al utilizar datos del microbioma (ARN 16S) y datos clínicos se podía detectar vaginosis bacteriana con

alta precisión (área bajo la curva (AUC del inglés, *Area Under a Curve*) 0.90) al compararla con métodos tradicionales (Beck and Foster 2015).

Naive Bayes. Es un clasificador lineal probabilístico basado en la aplicación del teorema de Bayes asumiendo independencia estadística entre las variables predictoras. En contraste con los modelos discriminativos como la regresión logística, el clasificador Naive Bayes es un modelo generativo ya que trata de entender las características independientes de cada clase para poder obtener asignar la probabilidad dada la presencia de ciertas características. Su objetivo es calcular la probabilidad de cada clase y seleccionar la que tenga mayor probabilidad. Para lograrlo se utiliza el teorema de Bayes: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ entendiendo que: $P(A|B)$ es la probabilidad A dado que B ya ocurrió; $P(B|A)$ es la probabilidad de B dado que A ya ocurrió; $P(A)$ es la probabilidad incondicional de que ocurra A, también se conoce como probabilidad previa; $P(B)$ es la probabilidad incondicional de que ocurra B. Utilizando las reglas Bayesianas se puede calcular la probabilidad condicional de pertenecer a cierta clase dado que ciertos valores en las variables independientes, para realizar clasificaciones acertadas. El algoritmo de Naïve Bayes es llamado naïve o vírgen porque realiza el supuesto de que hay dependencia entre las variables independientes. A pesar de que esto no ocurre en muchas situaciones de la vida real, es un algoritmo muy efectivo para resolver problemas complejos (Lakin 2021).

Algunos de los parámetros más importantes para modificar dentro del algoritmo se encuentran: *var_smoothing*: añadir un valor a la varianza de la distribución (por default = 1e-9). Este parámetro es importante porque comúnmente se usa una variante del algoritmo de Naive Bayes que permite extraer la distribución Gaussiana de los datos (media y varianza). Esta modificación en la distribución funciona como un filtro, dejando pasar solo a las muestras cercanas a la media dándoles un mayor peso. El parámetro *var_smoothing* ensancha la curva Gaussiana en el modelo, ocasionando que mayor muestras sean filtradas si se alejan más allá de la distribución media.

La aplicación de este algoritmo en datos del microbioma ha tenido importancia durante la clasificación taxonómica utilizando secuencias 16S ARN ribosomal (e.g. QUIME2) (Kuczynski et al. 2012).

Árboles de decisiones. Este formalismo permite resolver problemas de clasificación o regresión según una serie de preguntas y condiciones sobre los datos de entrada o que alimentan el modelo. La arquitectura del algoritmo de árboles de decisiones está compuesta por nodos internos, bordes o ramificaciones, hojas terminales y raíz. En la punta del árbol se encuentra la raíz que es el primer nodo de decisión, en cada nodo se elige una variable (en nuestro caso taxon bacteriano) y se realiza una pregunta o toma de decisión; los bordes o ramificaciones representan las respuestas a la pregunta; y las hojas terminales representan el resultado de salida sobre la etiqueta clasificada (figura, parte c). El objetivo del modelo es encontrar la arquitectura del modelo que más se acerque a la etiqueta real (Maimon and Lior 2014).

Para tomar una decisión en los nodos se toman en consideración todas las variables predictoras o independientes, y se selecciona la variable que produce mayor separación entre los datos observados. El objetivo es lograr grupos homogéneos o puros al realizar una toma de decisión sobre una variables, por ejemplo, separando en dos grupos a los sanos (clase 0) contra los enfermos (clase 1). La calidad de la separación puede ser medida mediante distintos criterios como el índice de Gini y la entropía, que miden la impureza o no homogeneidad entre las clases (Maimon and Lior 2014).

La entropía en los árboles de decisiones hace referencia a una medida de desorden que cuantifica la impureza (de 0 a 1) en un nodo. Por lo tanto, un valor alto (cercano a 1) de entropía contendría una mezcla de las etiquetas de la clase, por ejemplo, una mezcla entre los grupos sanos y enfermos. En cambio, un valor bajo (cercano a 0) de entropía refleja que el grupo está compuesto principalmente por una sola clase (e.g. enfermos o sanos). Al

realizar una decisión entre nodos puede ocasionar cierto grado de cambio en la entropía de los grupos conocido como ganancia de información. El objetivo de los árboles de decisiones es lograr la mayor ganancia de información, que se refleja en una reducción en la entropía entre nodos (Vidales 2019).

El índice de Gini como criterio de separación mide la probabilidad de hacer una incorrecta clasificación al escoger un elemento de forma aleatoria. Por consiguiente, un índice de bajo Gini tendría una menor probabilidad de una incorrecta clasificación. La finalidad del modelo sería encontrar las decisiones usando variables independientes que logren obtener un un valor Gini cercano al 0. Durante el proceso de entrenamiento de los árboles de decisión se pueden obtener diferentes resultados usando los distintos criterios de separación (índice de gini y entropía), sin embargo, su selección depende principalmente de cual ofrece mejor rendimiento en el modelo (Vidales 2019).

En general los árboles de decisiones son utilizados frecuentemente para resolver tareas de clasificación como métodos no paramétricos. Por ejemplo, Akira et al. en 2016 construyeron un modelo de árboles de decisiones utilizando datos de microbiota (ARN 16S) para clasificar obesos contra pacientes delgados. Este resultado resalta el potencial de usar modelos de ML para explicar problemas complejos como la relación microbiota-fenotipo del individuo (Andoh et al. 2016).

Random Forest. Es un método de ensamble (caracterizado por la construcción de varios clasificadores y su análisis estadístico promedio) basado en árboles de decisiones. Este algoritmo se forma por árboles de decisiones no correlacionados formados por la técnica de *bagging* (un tipo de método de ensamble). La técnica de *bagging* o bolseo se refiere a la utilización de datos de entrada obtenidos del conjunto de forma aleatoria para la construcción de *n* árboles de decisiones ocasionando baja correlación entre los árboles. La generación de árboles no correlacionados, y por lo tanto, diversos entre ellos otorga la

capacidad al modelo de aproximarse a las etiquetas reales. El resultado final del modelo es obtenido usando el promedio de las predicciones de cada árbol de decisiones (Breiman 2001).

Para obtener un adecuada precisión en el modelo se pueden ajustar los parámetros del algoritmo, entre los más importantes se encuentran los siguientes: *n_estimators:* cantidad de árboles que se crearán para entrenar al modelo (por defecto= 10); *max_depth*: profundida máxima del árbol (por defecto= ninguno*,* los nodos se expanden hasta que las hojas contienen la cantidad mínima de muestras); *min_samples_split*, número mínimo de muestras necesarias para dividir un nodo interno (por defecto= 2); *min_samples_leaf:* número mínimo de muestras requerido para estar en un nodo hoja (por defecto= 1)*; max_samples:* número máximo de muestras con las que se entrena cada árbol (por defecto= ninguna restricción); *max_features*: número máximo de variables independientes o características máximas cuando se realizan las decisiones (por defecto= ”*sqrt*”, la raíz cuadrada del número total de variables independientes) (Breiman 2001).

En conjunto la utilización de árboles no correlacionados permite realizar un modelo clasificador con adecuado rendimiento predictivo. En una revisión, (Zhou and Gallins 2019) et al. en 2019 comparon se comparó el AUC y precisión entre diferentes métodos de ML (incluyendo naive Bayes, regresión logística, aprendizaje profundo, Random Forest, etc.) usando 17 base de datos con diferentes fenotipos (incluyendo cirrosis hepática, EII, obesidad, distintas localizaciones anatómicas, etc.) El objetivo de los clasificadores era identificar el fenotipo utilizando datos del microbioma (16s ARN ribosomal). Se encontró que Random Forest tuvo el mejor resultado de precisión o estaba entre los métodos más competitivos en esta revisión con una AUC >0.85 para predecir el fenotipo del paciente (Zhou and Gallins 2019).

XGBoost (eXtreme Gradient Boosting). Es un método de ensamble tipo *boosting* basado en árboles de decisiones. Este algoritmo se realiza a partir de clasificadores débiles para formar un clasificador fuerte que se conoce como técnica *boosting* para mejorar la precisión del modelo. Los clasificadores débiles muestran una baja correlación con los valores reales.
Los árboles de decisión se crean calculando y minimizando el error del clasificador anterior, formando árboles de forma secuencial hasta lograr formar un clasificador con el rendimiento deseado. Una vez entrenado el modelo es capaz de realizar clasificaciones rápidas y precisas utilizando el clasificador fuerte formado.

Para lograr clasificadores más fuertes a partir del anterior, XGBoost utiliza el algoritmo gradiente descendiente para minimizar la distancia entre los valores predichos con los valores reales, conocida como función objetivo. Utilizando el algoritmo gradiente descendiente podemos encontrar el mínimo global de la función de costo de forma automática y eficiente. Otra característica importante es la aplicación de la poda (del inglés *pruning*) del árbol de decisión que permite reemplazar los nodos que no aportan a la mejora en la clasificación y evitando el sobre-ajuste del modelo. Para evitar el sobreajuste del modelo también se utilizan técnicas de regularización (como L1 o L2) que penalizan al modelo.

Algunos de los parámetros más importantes usados en el algoritmo XGBoost son los siguientes: *gamma*: número mínimo de reducción en la función de costo para dividir un nodo (por defecto=0); *max_depth*: profundidad máxima del árbol (por defecto=6); *max_delta_step*: el valor máximo que una hoja puede tener (por defecto=0, no hay restricción); *subsample*: porción de muestras del set para generar un clasificador (por defecto=1, se toma toda las muestras); *colsample*: porción de variables para generar cada clasificador (por defecto= 1, se toman todas las muestras); n_estimators: números de clasificadores que se van a generar (por defecto=100).

Un ejemplo del uso del algoritmo XGBoost fue cuando se usaron únicamente datos de secuenciación del ARN de célula única para identificar cáncer de mama metastásico usando. Este clasificador encontró una precisión alta con un AUC media de 0.82. Además, utilizando los valores de importancia de características se pudieron determinar que 6 genes (SQSTM1, GDF9, LINC01125, PTGS2, GVINP1, y TMEM64) podrían ser utilizados para predecir el estado de metástasis en el cáncer de mama (Li et al. 2022). Este artículo denota la importancia del uso de algoritmos de ML para identificar biomarcadores.

Redes Neuronales Artificiales. Es una rama de ML que utiliza una serie de capas neuronales compuesta por cierta cantidad de neuronas interconectadas obteniendo cierta información y dando un resultado de salida. El tamaño de la red es en función de la cantidad de neuronas en el modelo construido, siendo más complejo a medida que se agregan capas a la red. La unidad mínima del modelo se denomina perceptrón o neurona que está compuesto por la información de entrada a la neuronas con cierto peso. En conjunto el resultado de salida de las neuronas de una capa es sumadas y transformadas de forma no lineal dando el resultado activación (1) o no activación (0). El algoritmo llamado propagación hacia atrás (o *backpropagation*, en inglés) sirve para modificar los pesos en las redes neuronales y optimizar la predicción. Los pesos se modificarán la cantidad deseada hasta encontrar el error mínimo o diferencia entre los valores reales y la predicción del modelo.

De forma muy general, hay tres tipos de arquitecturas en el aprendizaje profundo: redes neuronales artificiales (perceptrón multicapas), redes neuronales convolucionales y autocodificadores (Lo and Marculescu 2019). De forma general para resolver problemas con datos tabulares se ha propuesto utilizar el perceptrón multicapas. La arquitectura de redes neuronales convolucionales ha sido ampliamente utilizada para resolver problemas de identificación y segmentación de imágenes. Los autocodificadores se ha descrito su utilidad en detección de imágenes, extracción de características, y sistemas de recomendación.

Métodos de ML	Definición - Características	Interpretabilidad	Ejemplos
Regresión logística	Es un método estadístico lineal que aprende a predecir el resultado de una variable dicotómica llamada *Y,* a partir de variables independientes (continuas o categóricas) llamadas *X*.	Método transparente: Se puede interpretar al leer el coeficiente de relación entre el resultado y las variables.	Diagnóstico vaginosis bacteriana usando datos del microbioma (16 RNA ribosomal) (Beck and Forest, 2015).
Naive Bayes	Calcula la probabilidad de una clase dado un conjunto de valores de características. Asume la independencia entre los atributos.	Método transparente: se puede interpretar al leer la contribución de cada variable en el resultado.	Clasificación taxonómica utilizando secuencias 16S ARN ribosomal (e.g. QUIME2).
Árboles de Decisiones	Resolver problemas de clasificación o regresión según una serie de preguntas y condiciones.	Método transparente: se puede interpretar al visualizar el árbol de decisión.	Técnicas de escrutinio para CRC basados en microbiota intestinal (Topçuoğlu et al. 2020).
Random Forest	Es un método de ensamble (combinación de múltiples clasificadores) basado en la generación de un conjunto de árboles de decisión no correlacionados que dependen de varias variables seleccionadas al azar. (técnica de bolseo) El resultado final del modelo es obtenido usando el promedio de las predicciones de los árboles de decisiones.	Método opaco: no se puede interpretar por sí solo el modelo, requiere herramientas *post-hoc* para el entendimiento de los resultados.	Identificación de pacientes con EII, cirrosis hepática, CRCl, y obesidad utilizando datos de 16 S ARN ribosomal. (Pasolli et al. 2016)
XGBoost	Es un método en ensamble (combinación de múltiples clasificadores) generando un clasificador fuerte creado reduciendo el error del modelo anterior hasta	Método opaco: no se puede interpretar por sí solo el modelo, requiere herramientas *post-hoc* para el entendimiento de	Identificación de cáncer de mama metastásico utilizando datos

	encontrar la predicción más alta. (técnica *boosting*)	los resultados.	RNA-seq de célula única. (Li et al. 2022)
Perceptrón multicapas (Redes neuronales artificiales)	Técnica de ML para extraer y transformar información utilizando múltiples capas de redes neuronales. Estas capas reciben información de las capas anteriores y se van refinando progresivamente. Se entrenan mediante algoritmos que minimizan los errores y mejoran la predicción.	Método opaco: no se puede interpretar por sí solo el modelo, requiere herramientas *post-hoc* para el entendimiento de los resultados.	Identificación de enfermedades (e.g. EII, CRC,cirrosis hepática y obesidad) utilizando datos metagenómicos del genoma completo. (Oh and Zhang 2020)

Tabla 1. Resumen de los métodos de aprendizaje de máquina, sus principales características, y algunos artículos característicos de su utilidad en medicina. CRC: cáncer colorrectal; EII: enfermedad inflamatoria intestinal.

3.4.2 Aprendizaje de máquina explicativo: modelos opacos y transparentes.

Un componente importante en la aplicación de los algoritmos de ML se encuentra en la explicabilidad del modelo. El objetivo es lograr que los modelos puedan ser comprendidos por el ser humano y así justificar el rendimiento, precisión y seguridad de las clasificaciones. En general no hay consenso para evaluar definir la capacidad de interpretación del modelo, sin embargo, existen ciertas propiedades que vuelven más comprensible al modelo y se puede explicar como se generó el resultado del modelo o poder simular los resultados del modelo. En general los algoritmos se pueden dividir en modelos opacos (cajas negras), y en modelos transparentes, de acuerdo a su capacidad para entenderlos. Se han clasificado como modelos transparentes, dada a su facilidad para entender los resultados, a los siguiente: regresión logística, árboles de decisiones, modelos bayesianos y los K vecinos más cercanos (*K-NN* del inglés, *K-nearest neighbors*). En cambios los modelos opacos, más difíciles de interpretar por sí solos, se encuentran los algoritmos más complejos como

Random Forest, XGBoost, y el aprendizaje profundo (exceptuando el perceptrón simple). Con el objetivo de explicar un modelo opaco, se han tratado de aplicar métodos *post-hoc* para que un modelo opaco se entienda como realiza sus clasificaciones, se pueden clasificar en modelos agnósticos y modelo específicos.

Entre estas herramientas *post-hoc* se pueden dividir en métodos agnósticos y los métodos basados en ejemplos. Los métodos diagnósticos se encargan de separar las explicaciones aparte del modelo de ML (no dependen de la arquitectura del modelo), utilizan los resultados y los datos de entrada para entender como se realizaron las clasificaciones. Entre los métodos agnósticos se encuentran el uso de explicaciones locales (e.g. algoritmo LIME), la explicación basada en la relevancia de la variables independientes (e.g. valores SHAP), y las explicaciones visuales (e.g. gráficas de dependencia parcial). Entre los modelos específicos o basados en ejemplos tienen el objetivo de seleccionar casos o muestras en particular para explicar el comportamiento del modelo como la obtención de un solo árbol de decisiones de un modelo Random Forest (Belle and Papantonis 2021).

Una de las técnicas más utilizadas es el uso de valores SHAP, que se basa en explicar el modelo de ML basado en la teoría del juego calculando la contribución de cada variable en el resultado de la clasificación. Los valores SHAP fueron introducidos en 1951 por Lloyd Shapley y sirve para explicar como distribuir el “pago” a cada “jugador” de forma justa por resolver una “tarea” en un juego cooperativo. Al extrapolarlo se trata de calcular la contribución (“pago”) de cada variable *X* (jugador) en el resultado del modelo (“tarea”) tomando en cuenta los diferentes ordenamiento de las variables *X.* Los posibles ordenamientos son importantes, ya que cada combinación entre los “jugadores” debe ser considerado para determinar la importancia de un solo “jugador” Por lo tanto, el valor SHAP se obtiene al promediar las contribuciones marginales de cada variable *X* tomando en cuenta todas los posibles ordenamientos de las variables *X*. La suma de los valores SHAP, incluyendo el valor base (la prevalencia de la clase positiva, en modelos de clasificación) es

igual al resultado del modelo o la probabilidad de pertenecer a una clase. La explicación local con la contribución de cada variable en el resultado lo podemos observar con las gráficas tipo *force plot* (figura 8 y figura 10). De la misma forma las explicación locales se pueden resumir en una gráfica para obtener la explicación global (figura 7, figura 9 y figura 11).

Estudiar la microbiota intestinal en los pacientes con DMT2 es fundamental para el entendimiento holístico de la enfermedad. Esta idea se reafirma al evaluar el impacto profundo que tiene la microbiota intestinal en la regulación del metabolismo, la modulación de la respuesta inmune y el mantenimiento de una mucosa intestinal saludable. Por otra parte, los métodos de ML se han propuesto para identificar nuevos tratamientos en los individuos con DMT2, debido a su gran capacidad para manejar una gran cantidad de datos e identificar las variables más relevantes asociadas al fenotipo en estudio. Por esas razones, la combinación de los perfiles de la microbiota intestinal y el uso de algoritmos de ML, podría permitir identificar a personas con riesgo de tener DMT2 y eventualmente sugerir líneas de tratamiento preventivo. Este acercamiento podría servir para señalar los taxones que se asocian con el estadio de la enfermedad, con la perspectiva de desarrollar nuevas intervenciones terapéuticas.

4 PREGUNTA DE INVESTIGACIÓN

¿Cuáles son los géneros bacterianos del microbioma intestinal caracterizados mediante 16s ARNr, en una población mexicana con distintos grados de avance de DMT2, que podrían estar asociados con el progreso de la enfermedad? De existir, ¿se podrán utilizar para clasificar correctamente a individuos con prediabetes o DMT2, de forma prospectiva?

5 PROPUESTA DE INVESTIGACIÓN

Partiendo de una cohorte de pacientes mexicanos, nuestra propuesta de investigación plantea utilizar diferentes métodos de aprendizaje de máquina supervisados, para proponer aquellas taxonomías y abundancias en el microbioma intestinal asociadas a distintos avances de la DMT2. El desarrollo de este objetivo permitirá asociar la composición del microbioma intestinal con el fenotipo de individuos sanos, con DMT2 y con alto riesgo de desarrollar DMT2. De forma relevante, nuestro estudio permitirá analizar las variables predictivas más representativas que hayan sido utilizadas para identificar correctamente a los pacientes al igual que su influencia en la participación del modelo predictivo.

6 HIPÓTESIS

Existe un grupo de bacterias, que permiten clasificar con precisión y de forma robusta a pacientes con DMT2 o con alto riesgo de desarrollar DMT2 (pre-DMT2). Estas bacterias podrían detentar propiedades de biomarcadores estando asociados al estado de salud o enfermedad del ser humano.

7 OBJETIVOS

7.1 General

El objetivo del estudio es la búsqueda de una posible asociación entre datos de microbiota intestinal y los diferentes estadios de la enfermedad en una cohorte de pacientes

mexicanos estratificados entre controles y pacientes con Pre-DMT2 y DMT2 sin tratamiento previo. Con este objetivo se pretende identificar posibles comunidades bacterianas en el microbioma intestinal que pudieran fungir como biomarcadores asociados con el progreso o riesgo de padecer diabetes tipo 2.

7.2 Específicos

- Comparar el rendimiento de seis distintos métodos de aprendizaje de máquina para clasificar a pacientes sanos contra enfermos utilizando datos del microbioma intestinal caracterizados mediante el gen marcador 16s ARNr. Los modelos de ML a evaluar son Regresión Logística, Naive Bayes, Árboles de Decisiones, Random Forest, XGBoost y Redes Neuronales Artificiales. En cada caso, se evaluará la precisión y área bajo la curva ROC (AUC, del inglés *Area Under the Receiver Operating Characteristics (ROC) Curve*) de cada modelo.
- Seleccionar el modelo con el mejor rendimiento en clasificación e identificar a los géneros bacterianos con mayor relevancia para la clasificación del fenotipo.
- Interpretar biológicamente los resultados obtenidos en términos de la DMT2, además, comparar los resultados obtenidos con aquellos reportados en otros estudios de otras poblaciones.

8 METODOLOGÍA DE INVESTIGACIÓN

8.1 Base de datos

La base de datos utilizada forma parte de un estudio previo (Diener et al. 2020), y se integra con datos de microbiota intestinal de 410 individuos del estado de Guanajuato,

México sin diagnóstico o tratamiento previo de DMT2. Los participantes fueron estratificados en: individuos sanos (n= 213), pacientes con pre-DMT2 (n= 150) y DMT2 (n= 47). Los sujetos de estudio fueron clasificados como individuos con pre-DMT2 al tener valores de glucosa plasmática en ayunas alterada (GAA) 100-125 mg/dl y/o con glucosa plasmática a las 2 horas alterada (TGA) con 140-199 mg/dl durante la prueba oral de tolerancia a la glucosa. Por otra parte, se clasificaron como individuos con DMT2 si tuvieron valores de glucosa en ayunas >126 mg/dl y/o con glucosa plasmática a las 2 horas >200 mg/dl.

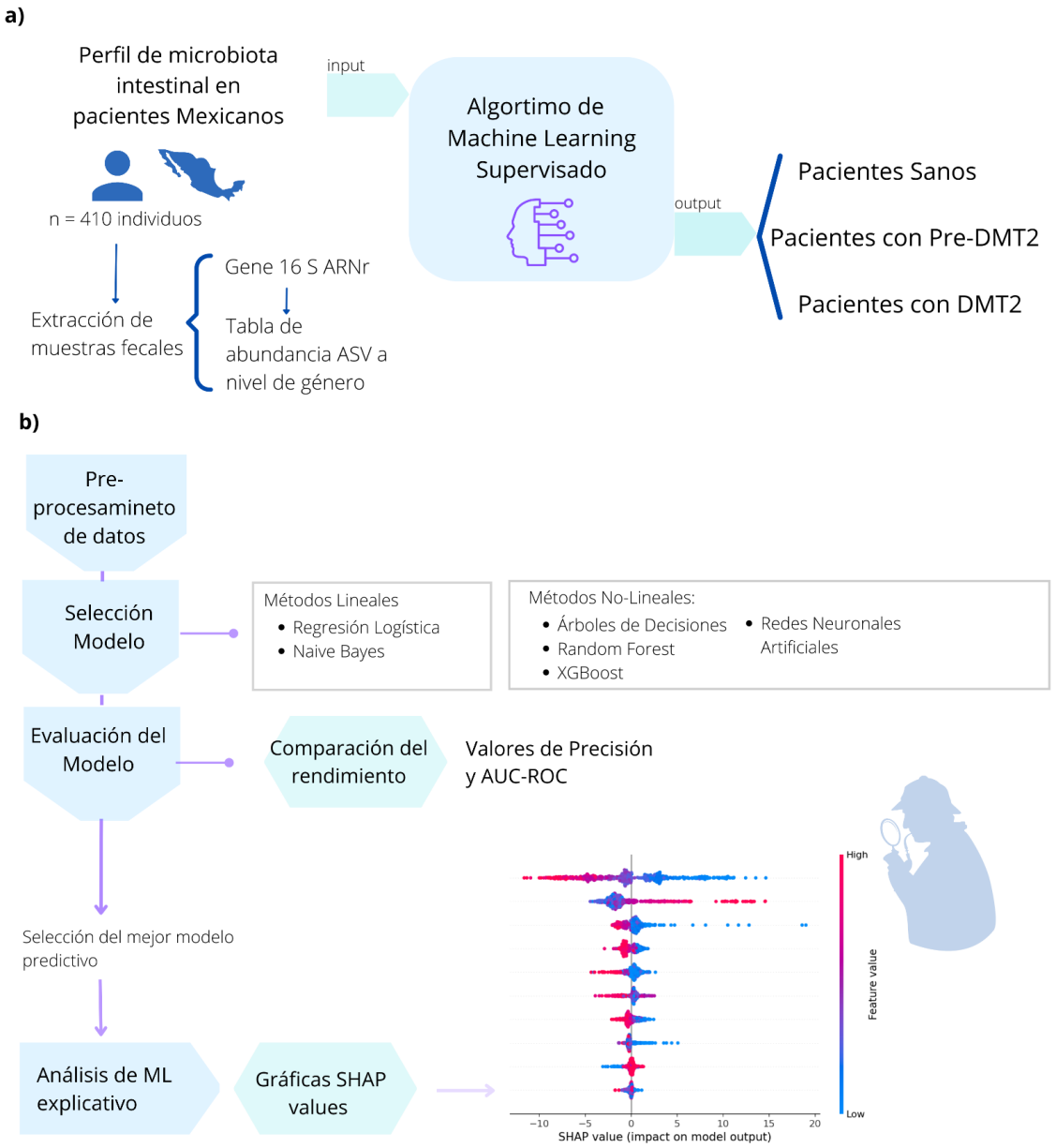

Figura 5. En la parte A, se observa el diseño de nuestra propuesta de investigación. En una cohorte de pacientes mexicanos del estado de Guanajuato con diferentes avances de DMT2 e individuos sanos, se obtuvo el perfil microbiano intestinal a partir de muestras fecales (16s ARNr). Mediante métodos de aprendizaje de máquina o ML (del inglés, Machine Learning) supervisado se realizaron predicciones del estadio del paciente: pre-DMT2, DMT2 o individuo sano. En la parte B, se muestra el flujo de trabajo de nuestro trabajo. Se realizó el pre-procesamiento de datos adecuado para los

distintos métodos de ML. Posteriormente se compararon el rendimiento predictivo de seis distintos métodos de ML, incluyendo métodos lineales y no-lineales. Finalmente, se seleccionó el método con mayor rendimiento predictivo y utilizando un acercamiento de ML explicativo (gráficas de valores SHAP) se identificaron las variables predictivas más importantes para el modelo, en este caso, géneros bacterianos.

8.2 Caracterización de la Microbiota intestinal por Secuenciación del gen 16S rRNA

En 410 muestras fecales se extrajo el ADN, y se realizó secuenciación de ampliaciones de la región hipervariable V4 del gen 16s ARNr. Esta región fue seleccionada debido a que la evidencia sugiere, en estudios predictivos, resulta en una mejor clasificación de la microbiota de muestras de enfermos contra muestras de sanos (Statnikov et al. 2013). Posteriormente, utilizando el flujo de trabajo del método *Divisive Amplicon Denoising Algorithm 2* (DADA2) (Callahan et al. 2016) se construyó una tabla de variantes de secuencia de amplicones (ASV, del inglés *Amplicon Sequence Variant*) que representa una versión de mayor resolución de la tabla de unidad taxonómica operativa (OTU, del inglés *Operational Taxonomic Unit*) producida por métodos tradicionales. La asignación taxonómica se realizó con la ayuda de la base de datos SILVA v132 (que proviene del latín *silva*, bosque) (Quast et al. 2013) Estos datos constituyen nuestro punto de partida para el desarrollo de esta tesis.

La base de datos final incluye 411 pacientes, incluyendo: 213 son pacientes sanos, 150 son pacientes con pre-DMT2, y 47 son pacientes con DMT2. De estas muestras se identificaron un total de 49,257,639 lecturas, de las cuales, se obtuvieron 17,059 ASVs. En general se obtuvo un promedio de 120,141 lecturas por muestra. Después de la asignación taxonómica se clasificaron 378 géneros, y posteriormente se filtraron obteniendo un total de 149 géneros, que constituyen el perfil taxonómico o perfil de abundancias de la base de datos final (tabla 2). La filtración se realizó con base en las siguientes reglas: se mantuvieron a los

géneros si tenían una abundancia media mayor a 10 lecturas en las muestras, y si mantuvieron a los géneros si al menos aparecían el taxón en el 10% de las muestras.

Índice	Actinomyces	Adlercreutzia	Akkermansia	Alistipes	Allisonella	Alloprevotella	...	Tyzzerella	Tyzzerella_3	Tyzzerella_4	Veillonella	Weissella	Estado
30099	34	39	42	139	0	0	...	15	0	0	73	114	0
30104	48	317	12421	362	0	83	...	0	536	0	110	79	0
30114	10	0	30110	5668	43	0	...	0	0	0	14	365	1
30170	0	0	2774	1030	12	0	...	0	0	0	357	0	1
30189	60	0	5387	888	0	0	...	0	1272	0	99	0	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...
31517	0	0	0	0	0	0	...	0	0	0	302	17	2
75007	45	0	22306	355	130	0	...	0	0	0	1578	0	0
75015	114	16	25	2097	0	300	...	0	16	0	385	731	1
75024	117	19	7973	2820	0	0	...	0	0	0	1864	11	1
75029	59	0	19	0	7	35	...	0	0	0	0	0	0

410 rows × 150 columns

Tabla 2. Tabla de abundancia a nivel de género. En las columnas se muestran los géneros bacterianos con un total de 149, y en las filas se encuentran los pacientes, que fueron clasificados en 3 estados (0: pacientes sanos, 1: pacientes pre-DMT2, 2: pacientes con DMT2).

8.3 Métodos - Aprendizaje de Máquina supervisado

Con la finalidad de identificar bacterias asociadas a cada estadio de la DMT2, se realizaron tres comparaciones *in silico* para probar las clasificaciones. En el primer análisis, clasificación 1 (C-1), se comparó los perfiles de abundancia y taxonomía bacteriana entre pacientes sanos contra pacientes con DMT2. En el segundo análisis, clasificación 2 (C-2), se comparó pacientes sanos contra pacientes con pre-DMT2. Finalmente, la clasificación 3 (C-3) consistió en una comparación multiclase con tres etiquetas: sanos, prediabéticos y diabéticos. Con la finalidad de entrenar los modelos de ML, se desarrolló un pipeline base en cada clasificación (figura 4) utilizando el lenguaje de programación Python 3.10.1 (Van

Rossum and Drake 2011), utilizando la paquetería Scikit-Learn, Pandas, Numpy y TensorFlow (Zaccone, Karim, and Menshawy 2017; Harris et al. 2020; McKinney 2010; Garreta and Moncecchi 2013).

En total se construyeron seis modelos de ML distintos y se evaluó su rendimiento predictivo en cada caso. Los métodos lineales utilizados, fueron: Regresión Logística Binaria y Naive Bayes. Los métodos no lineales utilizados, fueron: Árboles de Decisiones, Random Forest, XGBoost y Redes Neuronales Artificiales (con la arquitectura de Perceptrón-multicapa). Respecto a la arquitectura del perceptrón multicapas se construyó el modelo con una capa de entrada, dos capas profundas (32 neuronas en la primera capa y 16 en la segunda capa) y una capa de salida. La función de activación que se utilizó en las capas ocultas fue la función *Relu* (del inglés *Rectified linear unit*) que permite introducir la no-linealidad al modelo. La función de activación en la capa de salida utilizada fue la función sigmoide, que otorga una valor de predicción entre 0 y 1. En el caso de la clasificación multi-clase la función de activación en la capa de salida fue la función *softmax* que dió valores de probabilidad para varias clases (*Activation Functions in Deep Neural Networks* 2020). Para evitar el fenómeno de *overfitting* se utilizaron los siguientes métodos: se realizó un perceptrón multicapas sencillo (dos capas profundas), se añadieron dos capas *dropout* (con 0.5 o 50%), y se realizó la parada temprana del entrenamiento (paciencia = 10). El *dropout* es una técnica de regularización, donde se seleccionan cierto porcentaje de neuronas al azar para ser ignoradas durante el entrenamiento (en nuestro caso fue el 0.5 (50%) de las neuronas). La técnica *dropout* tiene el objetivo de una mejor generalización del modelo y evitar ser dependiente de las ciertas neuronas que pueden predisponer al fenómeno de *overfitting*. La parada temprana del entrenamiento también es una técnica de regularización, donde se detiene el entrenamiento del modelo cuando no se observa una mejora del rendimiento y así evitar el *overfitting o* sobre-entrenamiento (Salman and Liu 2019).

El flujo de trabajo del pipeline base para todos los modelos consistió en los siguientes pasos. Primero se dividió la base de datos de forma aleatoria en un conjunto de entrenamiento (80%) y un conjunto de prueba (20%). Después, se normalizaron (log2) en ambos conjuntos los valores del perfil microbioma intestinal (tablas de abundancia ASV). Este procedimiento se realizó únicamente en los métodos que requieren la normalización, tales como regresión logística, naive Bayes y en las redes neuronales artificiales. La normalización es un técnica que se utiliza cuando tiene rangos de valores muy altos o escalas diferentes, y permite la comparación entre las variables evitando que valores altos no tengan mayor importancia en el modelo únicamente por tener valores más altos. Posteriormente se entrenó de forma individual cada modelo utilizando todos los datos del conjunto de entrenamiento. Y después se evaluó la precisión del modelo utilizando los datos del conjunto de prueba.

La evaluación del modelo se realizó mediante las métricas de precisión y AUC-ROC. En el caso de la clasificación multiclase, clasificación C-3, se evaluó en conjunto con la métrica de precisión e índice de Kappa de Cohen. Para evaluar un modelo multi-clasificación se utiliza el coeficiente o índice de Kappa de Cohen que refleja la concordancia en los resultados que hay entre dos o más observadores o clasificadores tomando en cuenta el factor del azar. El rango de valores del coeficiente Kappa (κ) de Cohen se encuentra en -1 a +1 y un valor igual o menor a 0 indicaría que el resultado es exclusivo del azar. Si el resultado (κ) es entre 0.01-0.20 hay ligera concordancia entre clasificadores, si el resultado (κ) es entre 0.21-0.4 hay una aceptable concordancia entre clasificadores, si el resultado (κ) es entre 0.41-0.6 hay una moderada concordancia entre clasificadores, si el resultado (κ) es entre 0.6-0.8 hay buena concordancia entre clasificadores y si el resultado es mayor que 0.81 hay un una concordancia casi perfecta entre clasificadores (Kottner 2009).

Con la finalidad de evaluar el desempeño de cada modelo en función de los datos utilizados para entrenamiento, se utilizó la técnica de validación cruzada estratificada en cada clasificación en los seis modelos utilizados. Posteriormente, habiendo realizado este ensamble, se calculó la media y desviación estándar de la precisión y AUC-ROC obtenido en cada modelo entrenado. Los resultados se muestran en la figura 5, tabla 3, tabla 4, y tabla 5. La validación cruzada estratificada se realizó con un *K-fold* = 10, esto quiere decir que la base de datos en cada iteración se dividió en diez pliegues de forma aleatoria respetando el ratio de clases. De los diez pliegues generados, nueve sirvieron para entrar el modelo y el restante sirvió para probar la precisión del modelo. El objetivo es repetirlo diez veces con las distintas combinaciones de los pliegues. La media y la desviación estándar de los resultados se muestran en la figura 5 se obtuvieron al promediar las diez iteraciones descritas anteriormente.

Mediante la comparación de los distintos algoritmos pudimos seleccionar el modelo con mejor rendimiento predictivo y con menor dispersión de los resultados. Una vez que seleccionamos el mejor modelo de cada clasificación, entonces se procedió a realizar un análisis de interpretación *post-hoc*, esto utilizando los valores SHAP (del inglés SHapley Additive Explanations)(Lundberg et al. 2020). De forma relevante, estos parámetros nos permitieron identificar los géneros bacterianos de mayor importancia que utiliza el modelo de forma global durante la clasificación de individuos sanos o con algún avance de la enfermedad. Este último objetivo es uno de los propósitos centrales de este estudio, caracterizar las bacterias que potencialmente distinguen los estadios entre pre-diabetes y DMT2. Los conjuntos de datos generados y analizados durante la tesis, así como los pipelines utilizados, están disponibles en los repositorios de GitHub: [https://github.com/resendislab/Machine-Learning-Microbiome-T2D] A continuación discutimos en detalle cada una de las clasificaciones realizadas y las conclusiones derivadas en cada una de estas.

9 RESULTADOS. AVANCES DEL PROYECTO

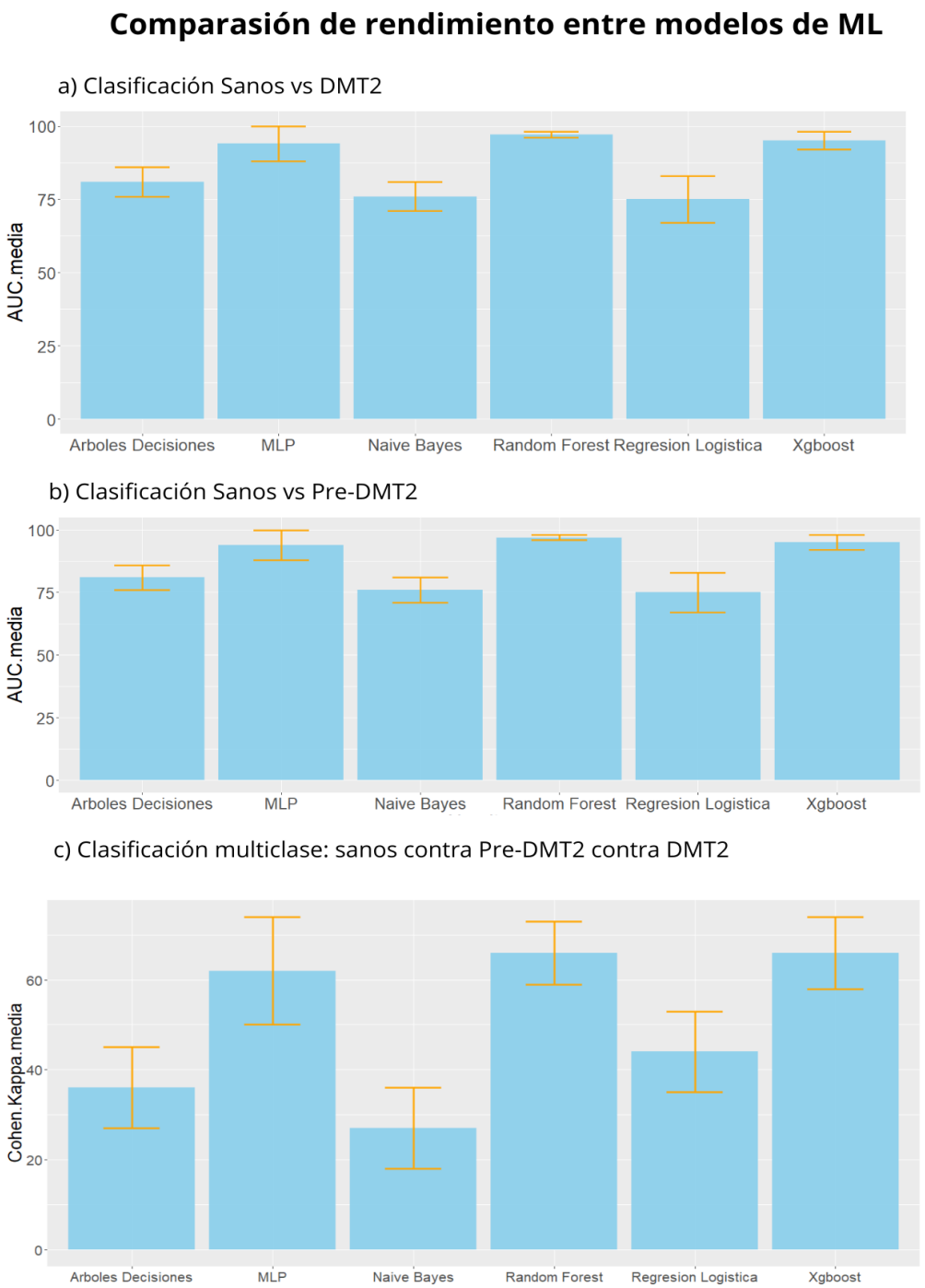

Figura 6. Comparamos seis algoritmos ML en tres clasificaciones: a, b y c. Se observa una gráfica de barras de error mostrando la media de los valores AUC-ROC con su respectiva desviación estándar

obtenida mediante la técnica de CV estratificada (CV=10) (amarillo). En el caso de la clasificación multiclase (parte C), la evaluamos con la media de Kappa de Cohen con su desviación estándar obtenido mediante la técnica de CV estratificada (CV=10).

Algoritmos de ML	Precisión	Sensibilidad	Especificidad	AUC	Precisión Media (DE, CV=10)	Precisión AUC (DE, CV=10)
Regresión Logística	0.85	0.73	0.96	0.85	0.70 (0.05)	0.75 (0.05)
Naive Bayes	0.74	0.63	0.83	0.75	0.69 (0.07)	0.76 (0.08)
Árboles de Decisiones	0.81	0.82	0.8	0.81	0.78 (0.06)	0.81 (0.05)
Random Forest	0.91	0.91	0.87	0.92	0.91 (0.04)	0.97 (0.01)
XG Boost	0.90	0.90	0.89	0.91	0.90 (0.05)	0.95 (0.03)
Redes Neuronales Artificiales (MLP)	0.88	0.89	0.96	0.91	0.94 (0.06)	0.94 (0.02)

Tabla 3. Clasificación 1 (Pacientes sanos vs Pacientes con DMT2). Comparamos el rendimiento de seis algoritmos de ML utilizando los valores de Precisión y AUC. Para obtener la desviación estándar se utilizó la técnica de CV estratificada con un *K fold* = 10. AUC: área bajo la curva, CV: validación cruzada (del inglés, Cross Validation) MLP: perceptrón multicapas (del inglés, Multilayer Perceptron), ML: Machine Learning

Algoritmos de ML	Precisión	Sensibilidad	Especificidad	AUC	Precisión Media (DE, CV=10)	Precisión AUC (DE, CV=10)
Regresión Logística	0.65	0.64	0.67	0.65	0.65 (0.09)	0.62 (0.07)
Naive Bayes	0.65	0.5	0.8	0.66	0.54 (0.05)	0.67 (0.06)
Árboles de Decisiones	0.61	0.51	0.72	0.62	0.61 (0.08)	0.64 (0.08)
Random Forest	0.80	0.85	0.76	0.80	0.74 (0.05)	0.82 (0.05)
XG Boost	0.68	0.67	0.68	0.73	0.69 (0.07)	0.75 (0.07)

Redes Neuronales Artificiales (MLP)	0.67	0.62	0.72	0.76	0.71 (0.07)	0.73 (0.08)

Tabla 4. Clasificación 2 (Pacientes sanos contra Pacientes con Pre-DMT2). Comparamos el rendimiento de seis algoritmos de ML utilizando los valores de Precisión y AUC. Para obtener la desviación estándar (DE) en nuestros resultados, utilizamos la técnica de validación cruzada estratificada (K Fold = 10) CV: validación cruzada (del inglés, Cross Validation) MLP: perceptrón multicapas (del inglés, Multilayer Perceptron), ML: Machine Learning

Algoritmos de ML	Precisión	Cohen Kappa	Precisión Media (DE, CV=10)	Cohen Kappa Media (DE, CV=10)
Regresión Logística	0.63	0.44	0.61 (0.03)	0.44 (0.09)
Naive Bayes	0.55	0.32	0.51 (0.1)	0.27 (0.09)
Árboles de Decisiones	0.61	0.42	0.58 (0.06)	0.36 (0.09)
Random Forest	0.76	0.64	0.77 (0.04)	0.66 (0.08)
XG Boost	0.77	0.66	0.77 (0.05)	0.66 (0.07)
Redes Neuronales Artificiales (MLP)	0.68	0.53	0.70 (0.09)	0.36 (0.09)

Tabla 5. Clasificación 3 (Pacientes sanos vs. Pacientes con Pre-DMT2 vs Pacientes con DMT2). Comparamos el rendimiento de seis algoritmos de ML utilizando los valores de puntuación de Precisión y Kappa de Cohen. Para obtener la desviación estándar (DE) en nuestros resultados, utilizamos la técnica de validación cruzada estratificada (K Fold = 10) CV: validación cruzada (del inglés, Cross Validation) MLP: perceptrón multicapas (del inglés, Multilayer Perceptron), ML: Machine Learning

9.1 Clasificación 1 (C-1): Personas sanas (n= 213) vs Personas con DMT2 (n= 47).

Los modelos con los mejores valores de precisión en la C-1 fueron: Random Forest (Precisión media = 0.91, desviación estándar (DE) 0.04) seguida de las redes neuronales artificiales (Precisión media = 0.94, DE 0.06). Los modelos con mejores valores AUC-ROC media en la C-1 fueron: Random Forest (AUC media = 0.97, DE 0.01) seguida de XGBoost (AUC media= 0.95, DE 0.03) (ver figura 5 y tabla 3).

Debido a su alto rendimiento predictivo se seleccionó al modelo Random Forest como el mejor modelo de la C-2 y se analizó mediante los valores SHAP (del inglés SHapley Additive Explanations) para identificar los géneros bacterianos más importantes para predecir individuos sanos (n= 213) contra individuos con DMT2 (n= 47). En la figura 6 se muestran, por orden de importancia, los 25 géneros bacterianos de mayor relevancia para clasificar los grupos de análisis. Además de la relevancia de cada bacteria para clasificar los grupos, en la figura 6 se muestra cómo los niveles de abundancia afectan las clasificaciones de ambos grupos. Así, los niveles altos (color rosa) de abundancia relativa de *Escherichia/Shigella, Slackia, Veillonella y Allisonella* contribuyen a seleccionar pacientes con DMT2 (valores SHAP positivos). En cambio, niveles altos (color rosa) de abundancia relativa de los géneros *Lachnospiracea UCG.004, Holdemanella, Ruminococcus 1, Ruminococcaceae_UCG 013, Ruminococcaceae_UCG 014, Ruminococcaceae_UCG 005, Enterococcus, Blautia, Lachnospiracea AC2044 group, y Anaerostipe*s ayudan a seleccionar los pacientes sanos (valores SHAP negativos).

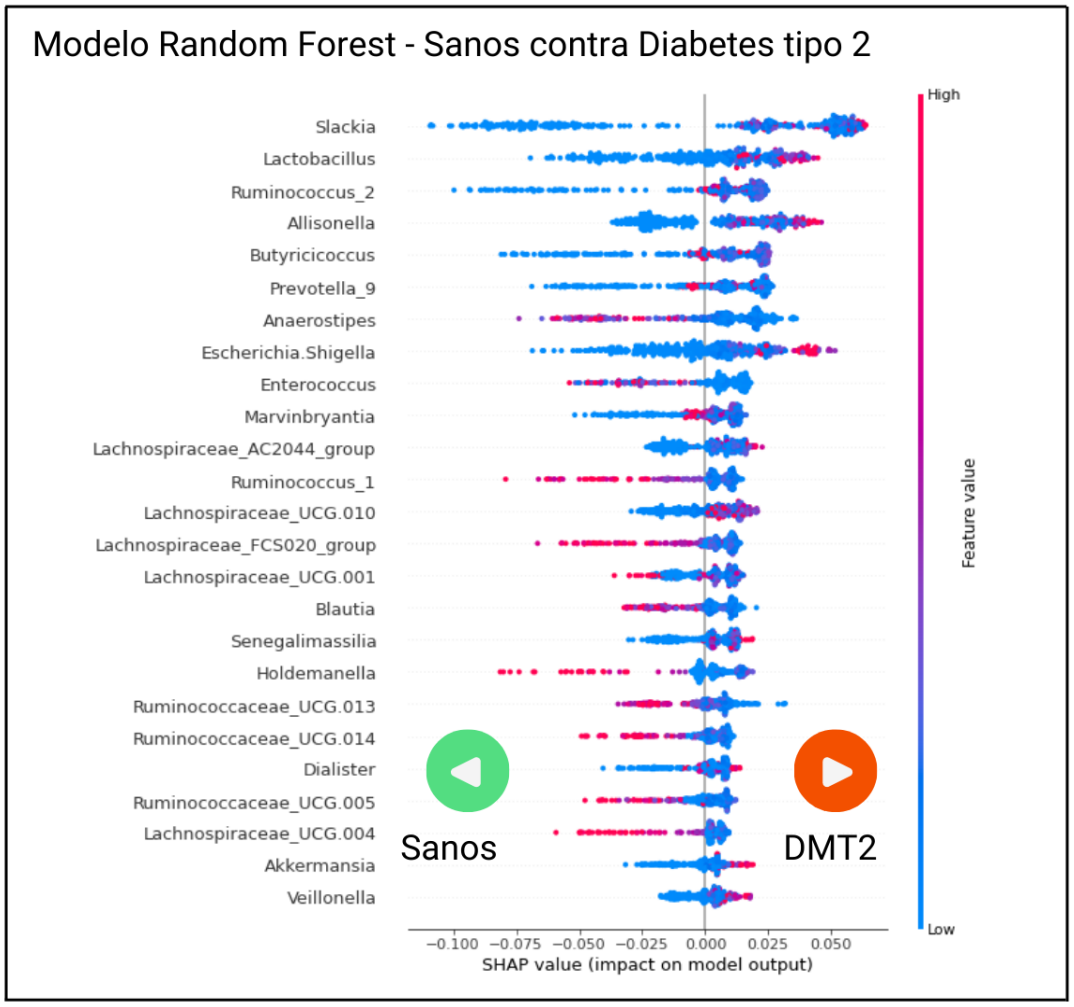

Figura 7. C-1: Pacientes sanos (n= 213) contra pacientes con DMT2 (n= 47). Mediante una gráfica de valores SHAP se muestra por orden de importancia los géneros bacterianos (mostrados en el eje de las Y) que tuvieron mayor responsabilidad en el modelo de Random Forest y el tipo de influencia que tienen en el modelo. Los valores SHAP positivos (mostrados en el eje X) son de utilidad para la clasificación de individuos con DMT2 y valores SHAP negativos (mostrados en el eje X) tiene mayor importancia para clasificar a pacientes sanos. Por código de colores se muestran los valores cuantitativos de la variable predictora, en este caso representa el valor de abundancia relativa del género bacteriano, siendo el color rosa para mostrar valores altos y el color azul para valores bajos de abundancia relativa.

Para la interpretación de casos específicos donde el modelo clasificó de forma incorrecta se generaron las gráficas de la figura 7. En estas gráficas se muestran 4 ejemplos de individuos clasificados de forma incorrecta, también se muestran los géneros de mayor importancia por cada caso, que empujaron el modelo hacia una clasificación como paciente con DMT2 o como individuo sano. Se seleccionaron dos ejemplos de casos de falsos negativos y dos ejemplos de casos de falsos positivos de forma aleatoria, obtenidos por el modelo Random Forest. En la parte A de la figura 7 se puede observar dos ejemplos donde el modelo Random Forest clasificó de forma incorrecta a individuo como paciente con DMT2, sin embargo, la etiqueta real de los individuos era ser sujeto sano (falso positivo). Podemos observar que tener valores altos de ciertos géneros como *Escherichia/Shigella, Prevotella_9, Lactobacillus, Ruminococcus_2* y *Slackia,* en este caso, ayudaron al modelo a clasificar como paciente con DMT2 de forma incorrecta (figura 7, parte A). En la parte B se pueden observar dos ejemplos en donde el modelo Random Forest clasificó de forma incorrecta como paciente sano a un individuo con DMT2 (falso negativo). Podemos observar, en la parte B de la figura 7, que los valores bajos de géneros como *Slackia, Rumiococcus_2 y Marvinbryantia* fueron de importancia para clasificar como un individuo sano comparado con DMT2. También se muestra que el cambio en ciertos géneros como el aumento de *Escherichia/Shigella* y *Prevotella_9*, y la disminución de *Anaerostipes* sirve para empujar el resultado de la clasificación hacia individuo con DMT2, sin embarazo, esta no es suficiente por lo que la clasificación final del modelo menciona que es un paciente sano (figura 7, parte B).

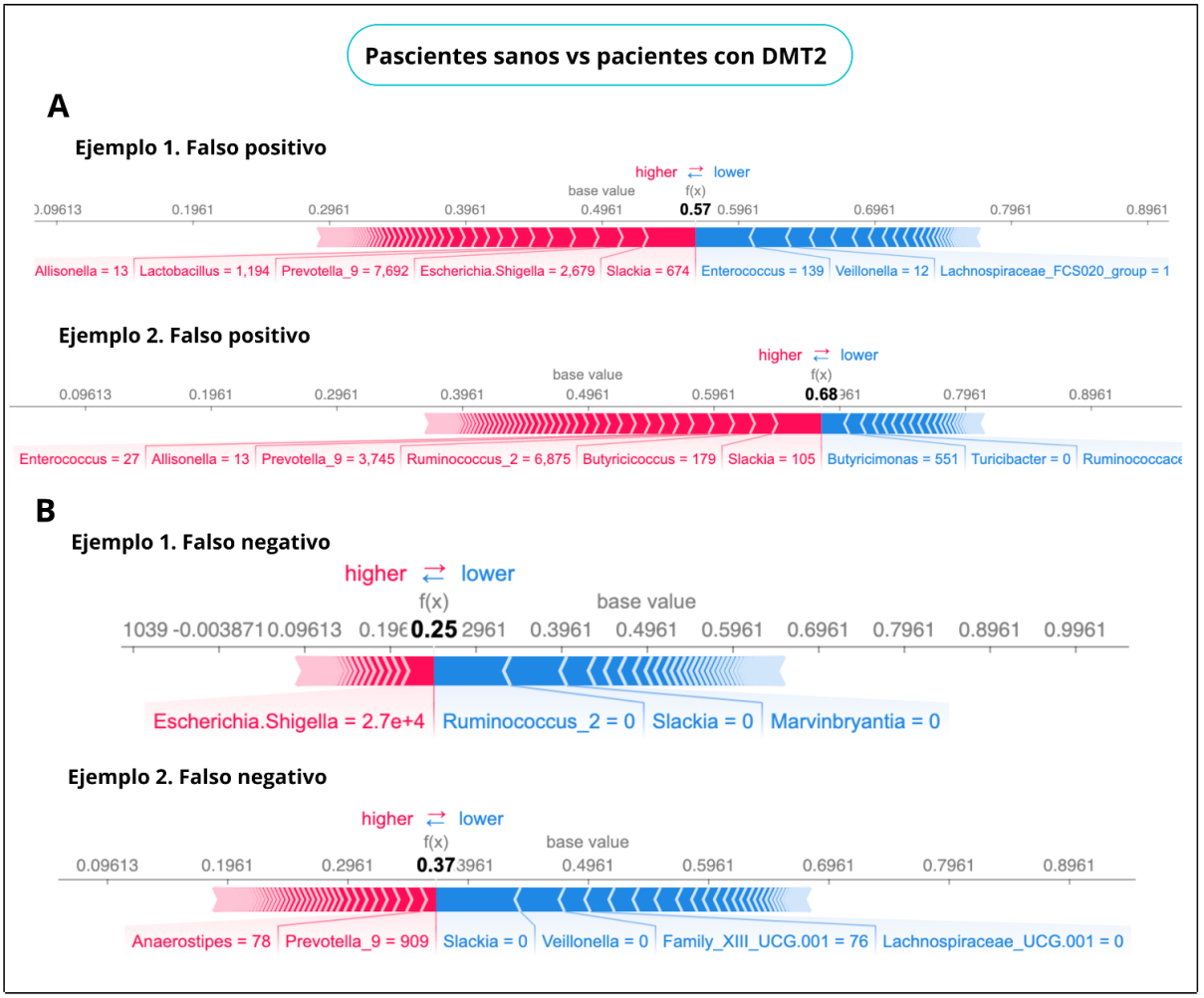

Figura 8. Gráficas de representativas de valores SHAP (tipo *Force plot)* usando dos ejemplos de falsos positivos en la parte A y dos ejemplos falsos negativos en la parte B. Obtenido los casos de forma aleatoria durante el modelo de Random Forest para clasificar a individuos con DMT2 contra individuos sanos (C-2: sanos vs DMT2). Por código de colores se puede visualizar si ciertos géneros fueron de mayor importancia para clasificar a un individuo como paciente con DMT2 (color rosa/rojo) o si fueron de mayor importancia para clasificar a un individuo como paciente sano (color azul). En las gráficas se muestran únicamente los géneros de mayor importancia para el modelo, y además se muestran sus valores de abundancia relativa que sirvieron como punto de decisión. El valor en color negritas (e.g. para el ejemplo 1 de la parte A es 0.57 y para el ejemplo 1 de la parte B es 0.25) muestra el resultado final del modelo para cada caso. Siendo valores cercanos a 1 da una

clasificación como paciente con DMT2 y valores bajos cercanos a 0 da una clasificación como paciente sano.

9.2 Clasificación 2 (C-2): Personas Sanas (n= 213) vs Personas con pre-DMT2 (n= 150)

En esta sección estudiaremos las diferencias entre personas sanas e individuos con pre-DMT2. Los modelos en la C-2 con los mejores valores de predicción en la clasificación fueron Random Forest (Precisión media= 0.74, DE 0.05) seguida de las redes neuronales artificiales (Precisión media= 0.71, DE 0.07). Los mejores modelos basados en la métrica AUC-ROC fueron Random Forest (AUC medio = 0.74, DE 0.05), seguida de XGBoost (AUC medio = 0.75, DE 0.07) (ver Figura 5 y tabla 4).

Con base a las métricas anteriores, se seleccionó el modelo Random Forest de la C-2 como el mejor clasificador en el rendimiento predictivo entre los seis modelos analizados. Mediante los valores SHAP, se identificaron a los géneros bacterianos más importantes útiles para predecir individuos con pre-DMT2 (n= 150) contra individuos sanos (n= 213). En la figura 8 se muestran, por orden jerárquico, los 25 géneros bacterianos más responsables del resultado del modelo, entre los que destacan: *Intestinibacter*, *Anaerostipes, Collinsella, Fusicatenibacter, Prevotella_9, Blautia, Escherichia/Shigella, y Granulicatella*.

Tal y como se muestra en la figura 8, los valores bajos de abundancia relativa (color azul) de *Intestibacter, Terrrisporobacter, Fusicatenibacter, Blautia, Allistipes, Rombustia, y Anaerostipes* ayudan a predecir los pacientes con pre-DMT2 comparado contra sanos. Por otra parte, los niveles altos de abundancia relativa (color rosa) de los géneros *Collinsella,*

Granulicatella, Veillonella, Slackia, Escherichia/Shigella y Solobacterium ayudan a distinguir a los pacientes con Pre-DMT2. Sin embargo, no es posible seleccionar un género único en nuestro caso para identificar con precisión a los individuos con la enfermedad; parece ser más significativo un conjunto de cambios específicos en el perfil de GM de los individuos con la enfermedad.

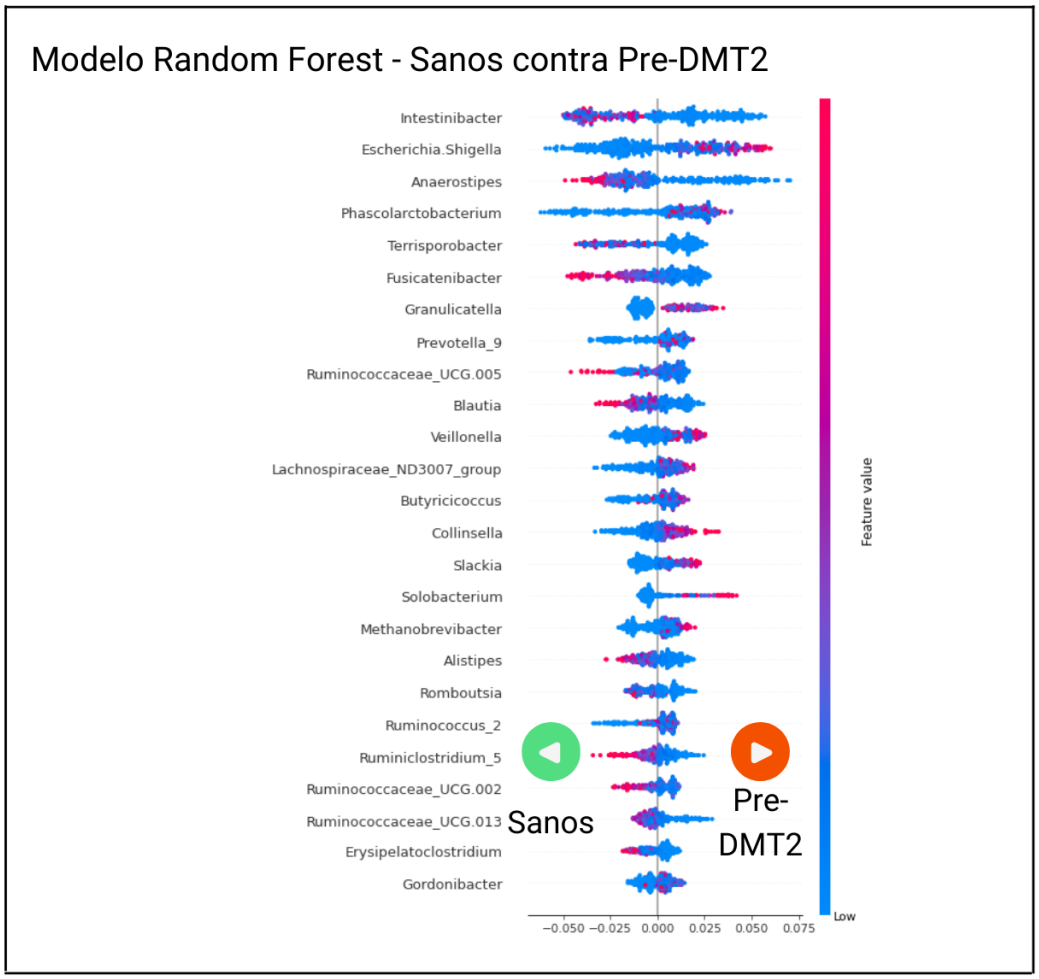

Figura 9. C-2: Pacientes sanos (n= 213) contra pacientes con pre-DMT2 (n= 150). Mediante una gráfica de valores SHAP se muestra por orden de importancia los géneros bacterianos (mostrado en el eje de las Y) que tuvieron mayor responsabilidad en el modelo de Random Forest y el tipo de peso que tienen en el modelo. Los valores SHAP positivos (mostrado en el eje de las X) son de utilidad

para la clasificación a individuos con DMT2, mientras que valores SHAP negativos son de utilidad para clasificar a pacientes sanos. Por código de colores se muestra la influencia que tienen los valores de variable predictiva en el resultado, siendo el color roja correspondiente a valores altos de abundancia relativa y el color azul correspondiente a valores bajos de abundancia relativa.

Para visualizar casos donde el modelo realizó una predicción incorrecta se generaron las siguientes gráficas en la figura 9. En estas gráficas se observan 4 ejemplos de individuos de predicciones incorrectas y se pueden interpretar para identificar cuales son los géneros de mayor importancia en la predicción de la etiqueta para cada caso. Se puede observar dos ejemplos donde el modelo Random Forest, clasificó de forma incorrecta a un paciente sano como paciente con pre-DMT2, conocido como falso positivo (figura 9, parte A). También dos ejemplos particulares donde el modelo Random Forest clasificó de forma incorrecta a pacientes con pre-DMT2 como paciente sano, conocido como falso negativo (figura 9, parte B). Respecto a la parte A en el ejemplo 1 podemos observar que valores relativamente bajos de géneros productores de AGCC como *Intestinibacter* y *Anaerostipes* empujan hacia una clasificación como pre-DMT2, al igual que valores altos de géneros que reflejan disbiosis intestinal como *Escherichia/shigella* empujaron el modelo hacia un clasificación como DMT2.

Otro ejemplo interesante es el que se muestra en la parte B de la figura 9, donde se puede observar que valores relativamente altos de abundancia relativa de ciertos géneros productores de AGCC como *Enterococcus* sirvió al modelo para clasificar como un individuo sano comparado contra un paciente con pre-DMT2.

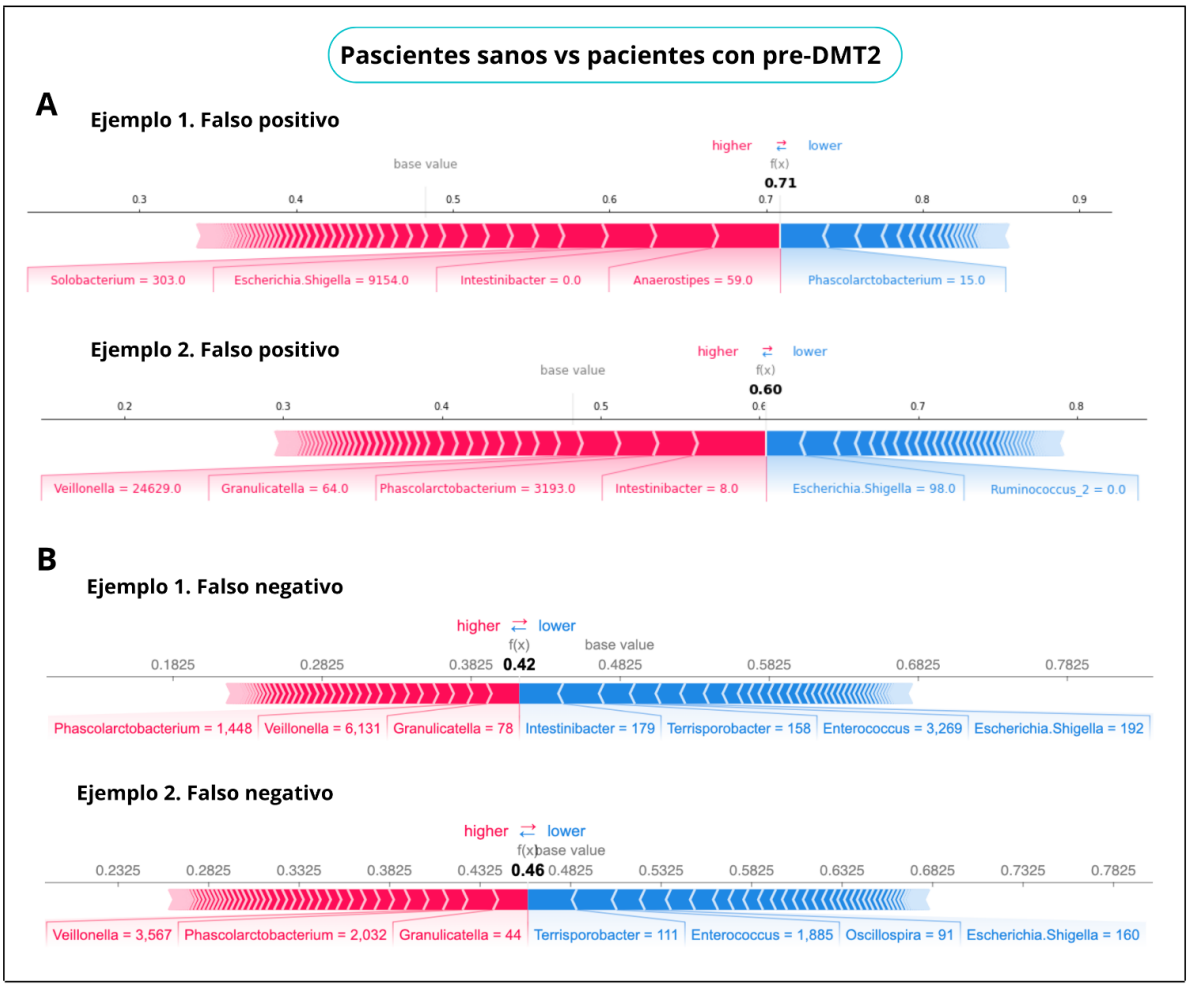

Figura 10. Gráficas de representativas de valores SHAP (tipo *Force plot)* usando dos ejemplos de falsos positivos en la parte A y dos ejemplos falsos negativos en la parte B. Obtenido los casos de forma aleatoria durante mediante el modelo de Random Forest para clasificar a individuos con pre-DMT2 contra individuos sanos (C-2: sanos vs pre-DMT2). Por código de colores se puede visualizar si ciertos géneros fueron de mayor importancia para clasificar a un individuo como paciente con pre-DMT2 (color rosa/rojo) o si fueron de mayor importancia para clasificar a un individuo como paciente sano (color azul). En las gráficas se muestran únicamente los géneros de mayor importancia para el modelo, y además se muestran sus valores de abundancia relativa que sirvieron como punto de decisión. El valor en color negritas (e.g. para el ejemplo 1 de la parte A es 0.57 y para el ejemplo 1 de la parte B es 0.25) muestra el resultado final del modelo para cada caso.

Siendo valores cercanos a 1 da una clasificación como paciente con pre-DMT2 y valores bajos cercano a 0 da una clasificación como paciente sano.

9.3 Clasificación 3 (C-3): Clasificación multiclase: Personas sanas (n= 213); Personas con pre-DMT2 (n= 150); personas con DMT2 (n = 47).

En el estudio de la C-3, los modelos con las mejores puntuaciones fueron: Random Forest (Precisión media= 0.77, DE 0.04) seguida de XGBoost (Precisión media= 0.77, DE 0.05). Con base, en la métrica de Kappa de Cohen los mejores modelos fueron: XGBoost (Kappa de Cohen =0.66, DE 0.07) seguida de Random Forest (Kappa de Cohen = 0.66, DE 0.08) (ver Figura 5 y tabla 3).

El modelo XGBoost de C-3 (multiclase) obtuvo el mejor rendimiento predictivo para la clasificación multiclase de individuos sanos (n= 213), Pre-DMT2 (n= 150) o DMT2 (n= 47). Analizamos este modelo de XGBoost mediante los valores SHAP, que nos permitió identificar los principales géneros bacterianos útiles para la clasificación multi-etiqueta: de individuos sanos, individuos con pre-DMT2, e individuos con DMT2. Entre los primeros 25 géneros bacterianos con mayor peso en la clasificación se muestran en la figura 10 por orden de importancia. Algunos de estos géneros son *Slackia, Ruminococcaceae_UCG.014, Faecalibacterium, Lactobacillus, Escherichia/Shigella, Allisonella, Prevotella_9, Anaerostipes, Veillonella, Akkermansia.* En la gráfica de la figura 10 se muestran los valores de SHAP promedio para las tres clasificaciones. En contraste con la figura 6 y 8, que muestran los valores SHAP positivos o negativos que ayudaban a mostrar la influencia de la abundancia relativa del género que tiene en el clasificador, aquí únicamente se indican los valores promedios del valor absoluto de SHAP. Está figura 10 nos permite observar los géneros de mayor importancia por orden jerárquico en un estudio multiclase.

Además, podemos observar cómo ciertos géneros bacterianos tienen un valor SHAP promedio más alto para clasificar a cierto estadio pre-DMT2 o DMT2. Por ejemplo, Slackia tiene valores promedios SHAP altos para predecir individuos sanos y diabéticos comparado con su valor promedio SHAP que es bajo para predecir individuos prediabéticos. También con *Lachnospiraceae UCG.10* tiene valores promedios SHAP altos para predecir diabéticos, al comparar con sus valores promedios de SHAP que son bajos para predecir individuos sanos y prediabéticos.

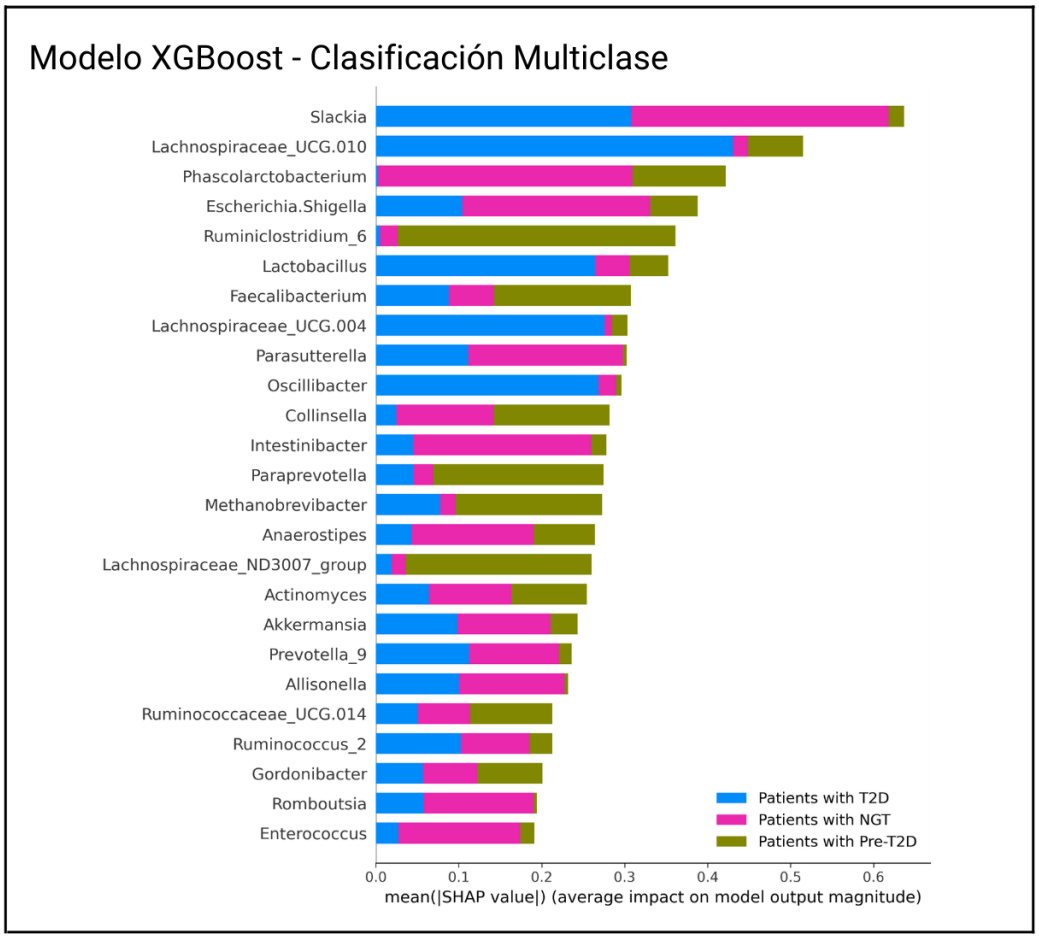

Figura 11. C-3: Clasificación multiclase: Personas Sanas (n= 213); Personas con pre-DMT2 (n= 150); personas con DMT2 (n = 47). En la parte A mediante una gráfica de valores SHAP se muestra por

orden de importancia los géneros bacterianos que tuvieron mayor responsabilidad en el modelo de XGBoost. Por código de colores se muestran las etiquetas del modelo, siendo la clase 0 (color rosa) como pacientes sanos, la clase 1 (color verde) como pacientes pre-diabéticos, la clase 2 (color azul) como pacientes diabéticos.

10 DISCUSIÓN DE RESULTADOS

La microbiota intestinal se ha propuesto como un factor emergente en la etiopatogenia de los individuos con DMT2, la cual se interrelaciona con los distintos factores de riesgo ambientales (como la dieta) y genéticos, mismos que perpetúan la afección sistémica de la enfermedad (Chakaroun, Massier, and Kovacs 2020). Sin embargo, estudiar la relación entre el huésped y la microbiota intestinal es compleja por lo que identificar los géneros característicos asociados a un estado prediabético o diabético continúa siendo un reto (Padron-Manrique et al., n.d.). Para contribuir a resolver esta cuestión, propusimos aplicar diferentes métodos de ML supervisado e identificar géneros bacterianos típicos de cada estadio de la DMT2 en población mexicana. En general, nuestro estudio nos permitió concluir que los algoritmos de ensamble basados en árboles de decisiones, como Random Forest y XGBoost, tuvieron el mejor rendimiento predictivo en nuestra cohorte. Realizando un análisis *post-hoc* de estos modelos, logramos identificar los taxones característicos y su impacto en los pacientes con DMT2 o pre-DMT2 al compararlo con su control negativo. Entre los géneros más importantes identificados por estos modelos se encuentran: *Escherichia/Shigella, Anaerostipes, Blautia, Intestinibacter, Collinsella.* En acuerdo con previos reportes, algunos estudios describen estos géneros con un rol potencial en la patogenia de la DMT2 en diferentes poblaciones (Padron-Manrique et al., n.d.; Esquivel-Herná et al. 2022; Balvers et al. 2021; Nie et al. 2021; X. Liu et al. 2021).

Como conclusión, en los siguientes párrafos se discuten los géneros más importantes identificados por los modelos para realizar las predicciones y cómo su disbiosis puede estar asociada con las alteraciones clínicas de la DMT2. De manera relevante, este estudio permite describir en detalle los cambios en la estructura de la microbiota intestinal y su asociación en una cohorte de pacientes mexicanos con DMT2 o pre-DMT2. Las alteraciones inducidas por la microbiota en el hospedero se pueden agrupar en tres fenómenos: 1) incremento en la permeabilidad intestinal, 2) reducción de los géneros productores de AGCC, y 3) alteración en la homeostasis intestinal y un incremento de géneros oportunistas (figura 11).

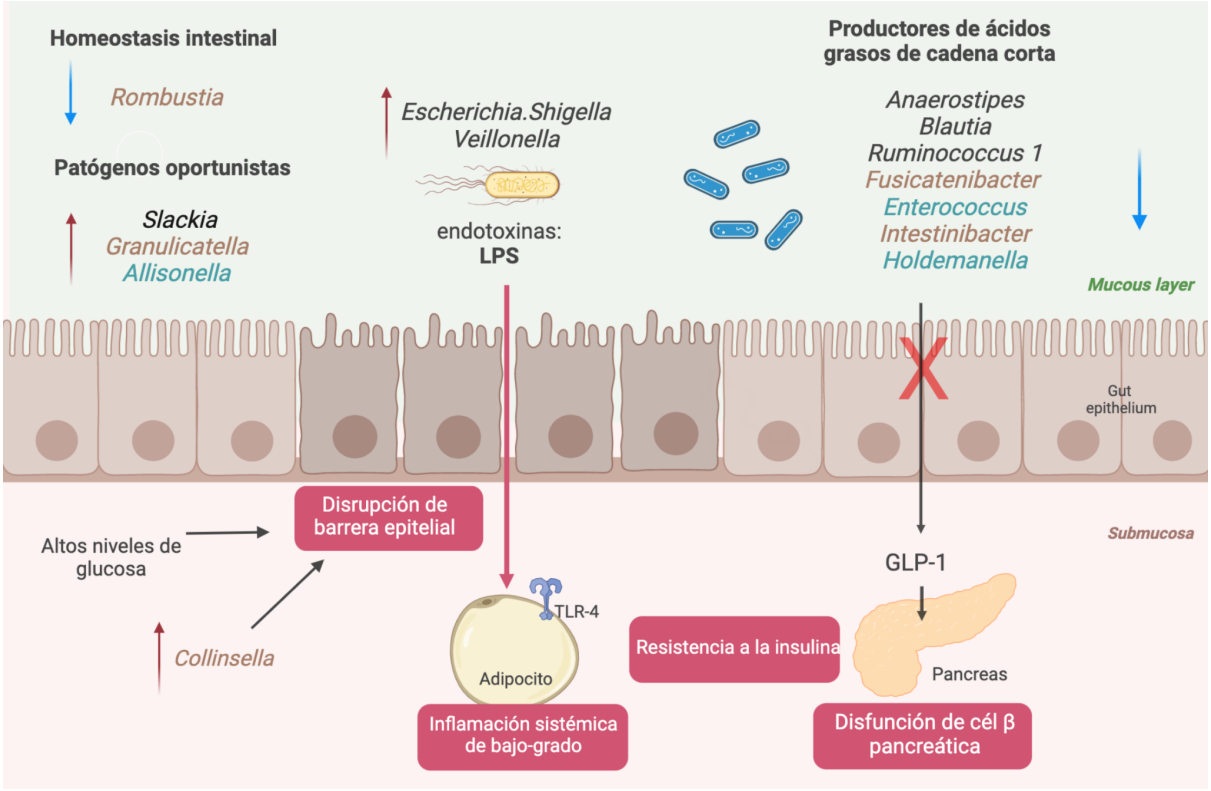

Figura 12. Se muestra un diagrama esquemático que ilustra los cambios en la microbiota intestinal asociado a la patogenia de los pacientes mexicanos con DMT2 o pre-DMT2. Hay tres principales componentes: 1) incremento en la permeabilidad intestinal 2) reducción de los géneros productores de AGCC 3) alteración en la homeostasis intestinal y un incremento de géneros oportunistas. En cada

componente se colocaron los géneros discriminatorios que sirvieron al modelo para clasificar a individuos con DMT2 (color aqua), pre-DMT2 (color café) o ambos (color negro) comparado contra sanos. AGCC: Ácidos grasos de cadena corta. DMT2: Diabetes Mellitus tipo 2.

Incremento de la permeabilidad intestinal.

Los niveles elevados de glucosa sanguínea han sido asociados a una pérdida de la integridad epitelial intestinal. Esta situación causa un aumento en la permeabilidad intestinal (Thaiss et al. 2018), y por lo tanto el paso de endotoxinas (como LPS (Lipopolisacáridos) desde el lumen intestinal hacia la circulación sistémica. A este fenómeno se le conoce como endotoxemia metabólica.

El paso de estas señales microbianas (incluyendo LPS, peptidoglicano y otros componentes) activan la respuesta inmune al ser reconocidas por sus receptores como el receptor tipo *toll* 4 (TLR4, *Toll-like receptor 4*, por sus siglas en inglés), TLR5, y TLR2. La interacción entre las señales microbianas y los receptores desencadena la expresión de citocinas proinflamatorias como la IL-4 (Interleucina-4), IL-6 (Interleucina-6), y el factor de necrosis tumoral alfa (por sus siglas en inglés, TNF-α). Este proceso es importante porque en los pacientes con DMT2 tiene característicamente un estado de inflamación sistémica de bajo grado. Esta inflamación sistémica se encuentra asociada a la progresión de la DMT2 y a sus complicaciones vasculares a largo plazo como IAM, EAP, y ACV (“Interactions between Gut Microbiota, Host Genetics and Diet Modulate the Predisposition to Obesity and Metabolic Syndrome” 2015; Torres-Leal et al. 2010).

Durante nuestro estudio encontramos que niveles altos de abundancia relativa en *Escherichia/Shigella* y *Veillonella* son necesarios para clasificar a individuos con DMT2 (C-1: sanos vs. DMT2) y a individuos con prediabetes (C-2: sanos vs. pre-DMT2). De acuerdo con la literatura, *Escherichia/Shigella y Veillonella* son géneros gram-negativos que contienen LPS en su pared celular. Se ha documentado en pacientes con DMT2 altos niveles de abundancia

relativa de estos géneros (Diener et al. 2020; Thingholm et al. 2019). Además, estos taxones han sido relacionados en condiciones no-diabéticas como síndrome de intestino irritable y enfermedad inflamatoria intestinal. Estos padecimientos tienen en común la inflamación local y sistémica asociada a una disbiosis intestinal que interviene en el progreso e inicio de la enfermedad (Chong et al. 2019; Kostic, Xavier, and Gevers 2014).

El aumento de la permeabilidad intestinal en pacientes con DMT2 también puede ser atribuido a otros factores como: consumo a largo plazo de alimentos procesados (dieta occidental), fármacos, consumo de alcohol y la disbiosis intestinal por sí misma (Bischoff et al. 2014) Por esta razón, se piensa que ciertos taxones específicos en la microbiota intestinal podrían afectar directamente la integridad epitelial. Algunos estudios señalan que *Collinsella* tiene un rol particular en este fenómeno. *Collinsella* interrumpe la barrera intestinal al disminuir la expresión de proteínas de unión tipo adherentes en los enterocitos (Bischoff et al. 2014; Chen et al. 2016). Consistente con esta observación, en nuestros análisis encontramos a *Collinsella* entre los primeros 25 géneros con mayor relevancia para clasificar a pacientes con prediabetes, y su un incremento en su abundancia relativa ayudaba a clasificar a estos pacientes comparado con control.

En conjunto, estos cambios señalan que la endotoxemia metabólica producto de múltiples factores del huésped (hiperglucemia) incluyendo la disbiosis intestinal podría contribuir a un estado inflamatorio sistémico crónico en pacientes con DMT2. Por lo que la medición de los metabolitos en el lumen intestinal y sistémica en pacientes mexicanos con pre-DMT2 y DMT2 ayudaría a entender las relaciones causales de esta asociación.

Reducción de los géneros productores de AGCC

En relación con la disbiosis intestinal, los pacientes con DMT2 tienen una disminución de los géneros productores de AGCC incluyendo butirato, propionato y acetato.

Además, una dieta tipo occidental (baja en fibra, rica en calorías provenientes de ácidos grasos saturados y azúcares) se asocia con una disminución de especies productoras de AGCC, principalmente butirato (Zhai et al. 2019). La baja producción de los AGCC ha sido asociada con alteraciones en la sensibilidad de la insulina y una inadecuada modulación del sistema inmune. Nosotros pudimos encontrar que los niveles bajos en géneros productoras de AGCC como *Anaerostipes, Blautia, Enterococcus, Intestinibacter* and *Fusicatenibacter*, ayudan a clasificar a pacientes con DMT2 o con pre-DMT2 comparado con pacientes sanos en cada caso. Además, estos géneros se encontraban entre las primeras 25 posiciones de mayor relevancia para el clasificador en la C-2: sanos vs. pre-DMT2 y la C-1: sanos vs. DMT2.

Durante nuestro estudio encontramos que niveles bajos de abundancia relativa en *Anaerostipes y Blautia,* se encontraba entre los primeros 25 de géneros más importantes para clasificar a individuos con DMT2 (C-1: sanos vs. DMT2) y a individuos con pre-DMT2 (C-2: sanos vs. pre-DMT2). Entre los primeros 25 de géneros más importantes, se encontró *Enterococcus* como taxón clave para clasificar a pacientes con DMT2 (C-1: sanos vs. DMT2). En la C-2: sanos vs. pre-DMT2 se encontró *Intestinibacter, Fusicatenibacter* dentro de los primeros 25 de géneros más importantes.

El butirato ha sido demostrado como un metabolito producido por la microbiota con efectos moduladores del sistema inmune y metabolismo del cuerpo humano. Se ha descrito que los AGCC disminuyen la producción de IL-6 (a nivel pancreático y tejido graso) en estudios pre-clínicos (Zhai et al. 2019; Fang et al. 2019) y clínicos (Larasati et al. 2019). Además, el butirato ayuda a promover la integridad epitelial, efectos anti-diabetogénicos, modula la sensibilidad de la insulina, y promueve un fenotipo normal de los colonocitos. Bajo este contexto, una microbiota intestinal “sana” mantiene una barrera epitelial en adecuado funcionamiento. Esto impide el paso de microorganismos y sus componentes que pueden activar al sistema inmune de una forma patológica. Por tal motivo, es esencial la

medición de metabolitos intraluminales en los pacientes con DMT2 para poder dilucidar la relación causal de este fenómeno.

Alteración en la homeostasis intestinal y un incremento de géneros oportunistas.

Podemos hipotetizar que ciertos cambios en la microbiota intestinal reflejan un nuevo estadio estable en el cuerpo humano, debido a una alteración de la homeostasis intestinal en pacientes con DMT2. Nosotros pudimos identificar ciertos géneros asociados con una alteración en la salud intestinal y ciertos patógenos oportunistas. Entre los que se encuentran: *Erysipelaclostridium, Escherichia/Shigella, Granulicatella, Allisonella, Rombustia,* y *Slackia.*

Se cree que este aumento de géneros oportunistas es resultado de la disbiosis intestinal y la pérdida de las bacterias productoras de AGCC (Qin et al. 2012). Es de relevancia mencionar que los pacientes con DMT2 en tratamiento con metformina se han observado un aumento de las especies productoras de butirato y estas podrían ayudar a restaurar el balance de la microbiota intestinal (Iulia-Suceveanu et al. 2019).

Identificamos entre los 25 mejores géneros bacterianos, a *Collinsella y Granulicatella*, que son de utilidad para clasificar a individuos con pre-DMT2 (C-2: sanos vs. DMT2). Estos géneros bacterianos se asocian a niveles elevados de Trimetilamina (TMA)(Fang et al. 2019; Y. Liu and Dai 2020). La TMA es producida por la microbiota intestinal mediante L-carnitina, colina y lecitina contenida en grandes cantidades en carnes rojas y alimentos grasos. En el hígado se produce TMAO (TMA oxidada) a partir de TMA, mediante la enzima FMO3 (del inglés flavin-containing monooxygenase 3) (Rajakovich et al. 2021) Altos niveles de este metabolito derivado de la microbiota intestinal, TMAO, tiene un rol en el patogénesis de la aterosclerosis al inducir una respuesta inflamatoria a nivel vascular, causando disfunción endotelial y alteración en el metabolismo del colesterol (Trøseid et al. 2015). Pacientes con

DMT2 y también con Pre-DMT2 tienen un significante riesgo de desarrollar enfermedades cardiovasculares. El metabolito TMAO puede estar relacionado como un factor determinante en la mortalidad de estos pacientes (Dambrova et al. 2016; Farhangi, Vajdi, and Asghari-Jafarabadi 2020) Se ha demostrado que altos niveles de TMAO se encuentra relacionado de forma positiva en la mortalidad y morbilidad de pacientes diabéticos y prediabéticos, sin embargo, es necesario seguir investigando para elucidar los posibles mecanismos de esta asociación.

Género	Función	Relación	DMT o pre-DMT2
Anaerostipes	Género productor de AGCC	Su disminución ayuda clasificar estado enfermo	DMT y pre-DMT2
Blautia	Género productor de AGCC	Su disminución ayuda clasificar estado enfermo	DMT y pre-DMT2
Collinsella	Relacionado con la disbiosis intestinal y el aumento permeabilidad intestinal	Su aumento ayuda clasificar estado enfermo	pre-DMT2
Enterococcus	Género productor de AGCC	Su disminución ayuda clasificar estado enfermo	DMT2
Escherichia/ Shigella	Relacionado con la disbiosis intestinal y la endotoxemia metabólica	Su aumento ayuda clasificar estado enfermo	DMT y pre-DMT2
Fusicatenibacter	Género productor de AGCC	Su disminución ayuda clasificar estado enfermo	pre-DMT2
Granulicatella	Identificado como patógeno oportunista en	Su aumento ayuda clasificar estado	pre-DMT2

	diversas situaciones clínicas	enfermo	
Intestinibacter	Género productor de AGCC	Su disminución ayuda clasificar estado enfermo	pre-DMT2
Rombustia	Se asociada a homeostasis intestinal en pacientes sanos	Su disminución ayuda clasificar estado enfermo	pre-DMT2
Slackia	Identificado como patógeno oportunista en pacientes con sepsis	Su aumento ayuda clasificar estado enfermo	DMT y pre-DMT2
Veillonella	Relacionado con la disbiosis intestinal y la endotoxemia metabólica	Su aumento ayuda clasificar estado enfermo	DMT y pre-DMT2

Tabla 6. Características importantes en el perfil taxonómico de pacientes con DMT2 y pre-DMT2. AGCC: ácidos grasos de cadena corta

Géneros indicadores de un individuo sano comparado con un paciente DMT2 o pre-DMT2

Durante las gráficas de las figuras 6 y 8, se pueden observar que ciertos géneros sirven para clasificar dependiendo su valor de abundancia relativa a un individuo sano comprado con un paciente con DMT2 (C-1) o un paciente con pre-DMT2 (C-2). Entre los géneros de relevancia algunos han sido descritos dentro de la literatura como predictores de una microbiota sana. Por ejemplo, la presencia e incremento de géneros productores de AGCC es fundamental para mantener un epitelio saludable al mantener el fenotipo normal de los colonocitos y además de ser moléculas moduladoras de sistema inmune vital para mantener una adecuada relación microbiota intestinal-mucosa intestinal en el ser humano (den Besten et al. 2013).

En la figura 6 se puede observar que el aumento de las abundancias relativas de géneros productores AGCC como A*naerostipes, Blautia, Holdemanella,* y *Enterococcus* sirvió

para el modelo Random Forest para clasificar a un individuo sano comparado con un paciente con DMT2. En acorde con esto, en la figura 8, podemos observar que el aumento de los géneros productores de AGCC como *Fusicatenibacterm, Blautia, y Anaerostipes* es de utilidad para clasificar a un individuo sano comparado contra paciente con pre-DMT2. *Blautia* ha sido identificado como género productor de AGCC, principalmente acetato, y con potenciales efectos probióticos (X. Liu et al. 2021). El acetato se une a sus receptores GPR41 y GPR43, y promueve el metabolismo de lípidos y glucosa disminuyendo las enfermedades relacionadas con la obesidad, como la DMT2 (Ang and Ding 2016).

En los individuos sanos se espera una ausencia o disminución de géneros oportunistas o potencialmente patógenos al compararlo con estados donde hay una disbiosis intestinal, como en pacientes con DMT2 (Bielka, Przezak, and Pawlik 2022). Es relevante destacar que la disminución de *Escherichia/Shigella* se encuentra entre los principales géneros para clasificar a un individuo sano comparado con un paciente con DMT2 (C-1) y también contra un paciente con pre-DMT2 (C-2). Entre otros géneros oportunistas que su nivel de abundancia relativa disminuida servía para clasificar a un individuo sano comparado con enfermo fueron: *Granulicatella, Allisonella, Megasphaera y Slackia.*

Falsos positivos y falsos negativos durante la evaluación del modelo.

El uso del perfil de microbiota intestinal fue de utilidad para clasificar a individuos con DMT2 e individuos con pre-DMT2 comparados con sanos con cierto grado de precisión (figura 5). Sin embargo, existen casos en donde el modelo realiza predicciones incorrectas sobre el fenotipo del paciente, conocidos como falsos positivos y falsos negativos. Para una interpretación específica de estos casos, se generaron la figura 7 y figura 9. En estas figuras se muestran por caso los géneros más importantes que ayudaron al modelo a realizar una

clasificación hacia la predicción de la etiqueta 1 (DMT2 o pre-DMT2) o hacia la etiqueta 0 (paciente sano).

En la figura 7 se muestran 4 ejemplos de clasificaciones incorrectas realizadas por el modelo Random Forest durante la C-1 (individuos sanos contra pacientes con DMT2). Dentro de estos ejemplos podemos observar que el modelo utiliza el aumento en la abundancia relativa de ciertos géneros (como *Escherichia/Shigella*), anteriormente descritos como característicos del estado diabético, para identificar a un individuo con DMT2 aunque esta predicción es incorrecta (figura 7, parte A). De igual forma, podemos observar que la disminución en la abundancia relativa de géneros característicos para el perfil diabético como *Veillonella y Slackia* conduce a el modelo hacia la predicción de un paciente sano de forma incorrecta en estos casos (figura 7, parte B).

En la figura 9 se muestran 4 ejemplos de clasificaciones incorrectas realizadas por el modelo Random Forest durante la C-2 (pacientes sanos contra pacientes con pre-DMT2). Es interesante mostrar que en estas gráficas la disminución de géneros productores de AGCC (como *Anaerostipes, Intestinibacter*) fue de utilidad para clasificar como paciente con pre-DMT2, aunque la etiqueta real de estos individuos era ser individuo sano (figura 9, parte B). Igualmente, los géneros como *Intestinibacter, Terrisporobacter, y Enterococcus* estuvieron como los taxones más importantes para clasificar a un individuo sano, aunque en estos casos de forma incorrecta.

En relación con la anterior, en ciertos casos, la utilización exclusiva del perfil microbioma intestinal para la predicción del fenotipo puede llevar a una mala clasificación. Por lo que la medición de las variables clínicas y laboratorio del paciente siguen siendo fundamentales para la valoración integral del individuo.

Finalmente, el trabajo de esta tesis favorece la idea de que la progresión de la patogenia de pacientes con DMT2 podría ser reflejada en los cambios dinámicos que presenta la microbiota intestinal. Por consiguiente, su entendimiento podría permitir una intervención en la historia natural de la enfermedad con la perspectiva de recibir intervenciones personalizadas. Notablemente, nuestro estudio apunta a que es difícil afirmar que un taxón en particular sea de utilidad para clasificar a pacientes sanos o enfermos. Todo lo contrario, una amplia caracterización del perfil microbioma intestinal o un conjunto de taxones microbianos es necesaria para encontrar valores óptimos en el posible tratamiento de pacientes, esto debido a la complejidad y heterogeneidad de los individuos que padecen DMT2.

12 CONCLUSIONES

En esta tesis, se ha descrito una asociación entre la microbiota intestinal y la progresión de la DMT2 en una cohorte de pacientes mexicanos. A futuro, los resultados podrían permitir el desarrollo de intervenciones y tratamientos oportunos basados en la microbiota que mejoren el pronóstico de la enfermedad. Por ejemplo, tratamientos nutricionales personalizados usando probióticos y la dieta con el enfoque de modificar la microbiota hacia el estado sano. Durante nuestra investigación se utilizaron distintos métodos de ML para identificar géneros bacterianos de gran relevancia para clasificar a pacientes prediabéticos y diabéticos contra controles en una cohorte de pacientes mexicanos. De los géneros de mayor importancia identificados, algunos han sido descritos en la literatura con un rol en la fisiopatología de la enfermedad, tales como *Escherichia/Shigella, Aanerostipes, Blautia, Intestinibacter, Collinsella, Prevotella.* Estos resultados basados en nuestra comunidad contribuyen a explorar la relación existente entre la microbiota y el desarrollo de biomarcadores que permiten identificar a personas con alto riesgo de desarrollar diabetes, con la perspectiva de recibir tratamientos preventivos y personalizados. Así, la utilización de estos modelos de ML podría ayudar a una

interpretación coherente de datos de alto rendimiento, identificación de biomarcadores potenciales y diseñar tratamientos personalizados para enfermedades metabólicas.

Durante nuestro estudio, con la finalidad de identificar bacterias asociadas a cada estadio de la enfermedad, se compararon seis métodos de clasificación: regresión logística, naive Bayes, árboles de decisiones, Random Forest, XGBoost, redes neuronales artificiales (perceptrón multicapas). El modelo con mayor precisión fue Random Forest para la C-1 (sanos vs. DMT2) y C-2 (sanos vs pre-DMT2). En el caso de la clasificación (C-3; sanos vs DMT2 vs pre-DMT2)) multiclase el mejor método fue XGBoost. Podemos concluir que en nuestro caso los métodos basados en árboles de decisiones son un acercamiento adecuado para modelar la asociación microbiota intestinal-DMT2. Esto concuerda con lo descrito ya para datos tabulares y de alta dimensionalidad, los modelos como XGBoost y Random Forest igualan o superan el rendimiento al compararlo con las redes neuronales artificiales (Uddin et al. 2019).

La utilización de algoritmos de aprendizaje profundo como las redes neuronales artificiales (perceptrón multicapas) han sido una alternativa atractiva debido a su poder predictivo. Entre nuestras clasificaciones, encontramos una AUC media= 0.94 (DE 0.06) para clasificar a pacientes con DMT2 vs sanos (C-1). Además, obtuvimos una AUC media= 0.71 (DE 0.07) para clasificar a pacientes con pre-DMT2 respecto de pacientes sanos (C-2). A pesar de que las AUC no son despreciables, concluimos que persisten algunos retos en su implementación. Uno de estos retos es utilizarla en bases de datos con poca cantidad de muestras. Se siguen realizando grandes avances en el área del aprendizaje profundo, y su creciente desarrollo seguirá siendo de interés en el área de la microbiota. De hecho, actualmente se puede interpretar los resultados de un modelo de aprendizaje profundo utilizando valores SHAP por lo que se vuelve una herramienta poderosa con la capacidad de interpretar los resultados.

Con estos resultados podemos concluir que los métodos basados en árboles de decisiones (Random Forest, XGBoost) igualan, y en ocasiones, con mayor desempeño que los modelos de aprendizaje profundo. Probablemente se deba a que los modelos de aprendizaje profundo son muy sensibles identificando patrones y se vuelven susceptibles al fenómeno de *overfitting*. A pesar de que existen varias herramientas para intentar solucionarlo, aún persiste este problema. En esta tesis, se aplicaron diferentes técnicas que pueden ser de utilidad para prevenir el *overfitting* como simplificar el modelo, añadir capas *dropout*, y el arresto temprano del entrenamiento del modelo. Consideramos que con una mayor cantidad de datos para entrenar el modelo de aprendizaje profundo permitiría encontrar mejores resultados de rendimiento alcanzando el potencial reportado en otros estudios.

Para entender mejor las implicaciones de nuestros resultados, consideramos que se debe complementar con nuevos estudios que permitan identificar confiablemente más allá del género bacteriano. Siendo la secuenciación tipo shotgun una metodología atractiva, que podría permitir estudiar las capacidades metabólicas de la microbiota intestinal además de la composición de la microbiota. Complementaria y necesariamente, estudios de metaboloma de la microbiota intestinal en pacientes diabéticos son datos que ayudarían enormemente a entender, validar, y mejorar varias de las conclusiones formuladas en esta tesis. En particular, esta última tecnología permitirá identificar y validar aquellas moléculas que produce la microbiota intestinal y ser indicadores del desarrollo de la DMT2. Es de relevancia mencionar que la metodología desarrollada en este trabajo, usando diferentes algoritmos de ML supervisado (como Random Forest o Redes Neuronales), se podría aplicar para ayudar a entender esta compleja relación entre el microbioma-metaboloma y su influencia en el estado de salud-enfermedad (Singh et al. 2019; Lee-Sarwar et al. 2020). Este acercamiento podría ayudar a la compresión de la influencia que tienen los metabolitos derivados de la microbiota en los pacientes con DMT2 mexicanos. Entre algunos de los

metabolitos de interés se encuentran: ácidos grasos de cadena corta, ácidos biliares secundarios, aminoácidos de cadena ramificada, aminoácidos derivados de indol y TMAO. Algunos de estos metabolitos han sido propuestos en esta tesis como determinantes de la DMT2. A pesar de que incluir datos del metaboloma en este estudio sería todavía un reto, indudablemente es una de las perspectivas a buscar en un futuro próximo (Mallick et al. 2017).

En resumen, el trabajo desarrollado en esta tesis nos permite aseverar que la microbiota intestinal como un componente clave para estudiar de forma integral a un paciente con DMT2. La disbiosis intestinal no únicamente es un reflejo del estado patológico del individuo, sino que también participa de manera activa favoreciendo el progreso de la enfermedad. Una modulación de la microbiota podría ser necesaria o favorable para regresar a una homeostasis intestinal acercando al paciente con DMT2 hacia un estado de salud global.

PERSPECTIVAS

El desarrollo de esta tesis nos permitió identificar un conjunto de taxones que son claves para identificar a individuos con DMT2 y pre-DMT2 comparado con sanos. Esto se logró utilizando modelos de ML que identificaban con precisión a los diferentes estadios. Posteriormente mediante métodos *post-hoc* pudimos interpretar a los clasificadores y la influencia de los principales géneros bacterianos. Con la misma metodología desarrollada en esta tesis se podría identificar en una cohorte externa de pacientes mexicanos a individuos con alto riesgo de desarrollar DMT2. También se podría entender cuáles son los géneros característicos que influyen hacia un estadio de salud: sano o enfermo. Estos conjunto de

taxones de cada individuo podrían representar la firma única de su microbiota intestinal asociada a la enfermedad.

Consideramos necesario validar este conjunto de taxones característicos con una metodología experimental que se encuentre dirigida hacia estos géneros bacterianos descritos en la tesis. Esto es fundamental ya que se han descrito muchos hallazgos diferentes a nivel de género en estudios de la microbiota. Se cree que es debido a las diferentes especies o cepas que se capturan en el mismo género bacteriano podrían tener una función distinta. Además, es fundamental para determinar la contribución causal de los taxones en la relación microbiota intestinal-diabetes.

Bibliografía

Aagaard, Luna, and Versalovic. n.d. “The Human Microbiome of Local Body Sites and Their Unique Biology.” *And Bennett’s Principles and Practice of …*.

Activation Functions in Deep Neural Networks. 2020. North-West University, Potchefstroom Campus.

Alegre-Díaz, Jesus, William Herrington, Malaquías López-Cervantes, Louisa Gnatiuc, Raul Ramirez, Michael Hill, Colin Baigent, et al. 2016. “Diabetes and Cause-Specific Mortality in Mexico City.” *The New England Journal of Medicine* 375 (20): 1961–71.

Al Nabhani, Ziad, Sophie Dulauroy, Rute Marques, Clara Cousu, Shahed Al Bounny, François Déjardin, Tim Sparwasser, Marion Bérard, Nadine Cerf-Bensussan, and Gérard Eberl. 2019. “A Weaning Reaction to Microbiota Is Required for Resistance to Immunopathologies in the Adult.” *Immunity* 50 (5): 1276–88.e5.

Al Nabhani, Ziad, and Gérard Eberl. 2020. “Imprinting of the Immune System by the Microbiota Early in Life.” *Mucosal Immunology* 13 (2): 183–89.

American Diabetes Association. 2013. “Diagnosis and Classification of Diabetes Mellitus.” *Diabetes Care* 36 Suppl 1 (January): S67–74.

———. 2021a. “9. Pharmacologic Approaches to Glycemic Treatment: Standards of Medical Care in Diabetes-2021.” *Diabetes Care* 44 (Suppl 1): S111–24.

———. 2021b. “10. Cardiovascular Disease and Risk Management:” *Diabetes Care* 44 (Suppl 1): S125–50.

Andoh, Akira, Atsushi Nishida, Kenichiro Takahashi, Osamu Inatomi, Hirotsugu Imaeda, Shigeki Bamba, Katsuyuki Kito, Mitsushige Sugimoto, and Toshio Kobayashi. 2016. “Comparison of the Gut Microbial Community between Obese and Lean Peoples Using 16S Gene Sequencing in a Japanese Population.” *Journal of Clinical Biochemistry and Nutrition* 59 (1): 65–70.

Ang, Zhiwei, and Jeak Ling Ding. 2016. “GPR41 and GPR43 in Obesity and Inflammation - Protective or Causative?” *Frontiers in Immunology* 7 (February): 28.

Atlas, Diabetes, and Others. 2015. “International Diabetes Federation.” *IDF Diabetes Atlas, 7th Edn. Brussels, Belgium: International Diabetes Federation* 33. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.678.2328&rep=rep1&type=pdf.

Balvers, Manon, Mélanie Deschasaux, Bert-Jan van den Born, Koos Zwinderman, Max Nieuwdorp, and Evgeni Levin. 2021. “Analyzing Type 2 Diabetes Associations with the Gut Microbiome in Individuals from Two Ethnic Backgrounds Living in the Same Geographic Area.” *Nutrients* 13 (9). https://doi.org/10.3390/nu13093289.

Barone, Mark T. U., and Luiz Menna-Barreto. 2011. “Diabetes and Sleep: A Complex Cause-and-Effect Relationship.” *Diabetes Research and Clinical Practice* 91 (2): 129–37.

Barquera, S., I. Campos, and J. A. Rivera. 2013. “Mexico Attempts to Tackle Obesity: The Process, Results, Push Backs and Future Challenges.” *Obesity Reviews: An Official Journal of the International Association for the Study of Obesity* 14 Suppl 2 (November): 69–78.

Barquera, Simon, Ismael Campos-Nonato, Carlos Aguilar-Salinas, Ruy Lopez-Ridaura, Armando Arredondo, and Juan Rivera-Dommarco. 2013. “Diabetes in Mexico: Cost and

Management of Diabetes and Its Complications and Challenges for Health Policy." *Globalization and Health* 9 (February): 3.
Beck, Daniel, and James A. Foster. 2015. "Machine Learning Classifiers Provide Insight into the Relationship between Microbial Communities and Bacterial Vaginosis." *BioData Mining* 8 (August): 23.
Belle, Vaishak, and Ioannis Papantonis. 2021. "Principles and Practice of Explainable Machine Learning." *Frontiers in Big Data* 4 (July): 688969.
Besten, Gijs den, Aycha Bleeker, Albert Gerding, Karen van Eunen, Rick Havinga, Theo H. van Dijk, Maaike H. Oosterveer, et al. 2015. "Short-Chain Fatty Acids Protect Against High-Fat Diet-Induced Obesity via a PPARγ-Dependent Switch From Lipogenesis to Fat Oxidation." *Diabetes* 64 (7): 2398–2408.
Besten, Gijs den, Karen van Eunen, Albert K. Groen, Koen Venema, Dirk-Jan Reijngoud, and Barbara M. Bakker. 2013. "The Role of Short-Chain Fatty Acids in the Interplay between Diet, Gut Microbiota, and Host Energy Metabolism." *Journal of Lipid Research* 54 (9): 2325–40.
Bhattacharya, Prasanta K., and Aakash Roy. 2016. "Primary Prevention of Diabetes Mellitus: Current Strategies and Future Trends." *Italian Journal of Medicine*. https://doi.org/10.4081/itjm.2016.634.
Bielka, Weronika, Agnieszka Przezak, and Andrzej Pawlik. 2022. "The Role of the Gut Microbiota in the Pathogenesis of Diabetes." *International Journal of Molecular Sciences* 23 (1). https://doi.org/10.3390/ijms23010480.
BIRTH-GENE (BIG) Study Working Group, Tao Huang, Tiange Wang, Yan Zheng, Christina Ellervik, Xiang Li, Meng Gao, et al. 2019. "Association of Birth Weight With Type 2 Diabetes and Glycemic Traits: A Mendelian Randomization Study." *JAMA Network Open* 2 (9): e1910915.
Bischoff, Stephan C., Giovanni Barbara, Wim Buurman, Theo Ockhuizen, Jörg-Dieter Schulzke, Matteo Serino, Herbert Tilg, Alastair Watson, and Jerry M. Wells. 2014. "Intestinal Permeability--a New Target for Disease Prevention and Therapy." *BMC Gastroenterology* 14 (November): 189.
Bisong, Ekaba. 2019. "Logistic Regression." *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, 243–50.
Borenstein, Elhanan, and Efrat Muller. 2021. "Faculty Opinions Recommendation of Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment." *Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature*. https://doi.org/10.3410/f.739778223.793587742.
Bouskra, Djahida, Christophe Brézillon, Marion Bérard, Catherine Werts, Rosa Varona, Ivo Gomperts Boneca, and Gérard Eberl. 2008. "Lymphoid Tissue Genesis Induced by Commensals through NOD1 Regulates Intestinal Homeostasis." *Nature* 456 (7221): 507–10.
Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.
Bullock, Ann, and Karen Sheff. 2017. "Incidence Trends of Type 1 and Type 2 Diabetes among Youths, 2002-2012." *The New England Journal of Medicine*.
Cai, Xiaoyan, Yunlong Zhang, Meijun Li, Jason Hy Wu, Linlin Mai, Jun Li, Yu Yang, Yunzhao Hu, and Yuli Huang. 2020. "Association between Prediabetes and Risk of All Cause Mortality and Cardiovascular Disease: Updated Meta-Analysis." *BMJ* 370 (July): m2297.
Callahan, Benjamin J., Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A.

Johnson, and Susan P. Holmes. 2016. “DADA2: High-Resolution Sample Inference from Illumina Amplicon Data.” *Nature Methods* 13 (7): 581–83.
Carvalho, Bruno Melo, and Mario Jose Abdalla Saad. 2013. “Influence of Gut Microbiota on Subclinical Inflammation and Insulin Resistance.” *Mediators of Inflammation* 2013 (June): 986734.
Chakaroun, Rima M., Lucas Massier, and Peter Kovacs. 2020. “Gut Microbiome, Intestinal Permeability, and Tissue Bacteria in Metabolic Disease: Perpetrators or Bystanders?” *Nutrients* 12 (4). https://doi.org/10.3390/nu12041082.
Chatterjee, Sudesna, Melanie J. Davies, Simon Heller, Jane Speight, Frank J. Snoek, and Kamlesh Khunti. 2018. “Diabetes Structured Self-Management Education Programmes: A Narrative Review and Current Innovations.” *The Lancet. Diabetes & Endocrinology* 6 (2): 130–42.
Chatterjee, Sudesna, Kamlesh Khunti, and Melanie J. Davies. 2017. “Type 2 Diabetes.” *The Lancet*. https://doi.org/10.1016/s0140-6736(17)30058-2.
Chen, Jun, Kerry Wright, John M. Davis, Patricio Jeraldo, Eric V. Marietta, Joseph Murray, Heidi Nelson, Eric L. Matteson, and Veena Taneja. 2016. “An Expansion of Rare Lineage Intestinal Microbes Characterizes Rheumatoid Arthritis.” *Genome Medicine* 8 (1): 43.
Chong, Pei Pei, Voon Kin Chin, Chung Yeng Looi, Won Fen Wong, Priya Madhavan, and Voon Chen Yong. 2019. “The Microbiome and Irritable Bowel Syndrome - A Review on the Pathophysiology, Current Research and Future Therapy.” *Frontiers in Microbiology* 10 (June): 1136.
Colchero, M. Arantxa, Mariana Molina, and Carlos M. Guerrero-López. 2017. “After Mexico Implemented a Tax, Purchases of Sugar-Sweetened Beverages Decreased and Water Increased: Difference by Place of Residence, Household Composition, and Income Level.” *The Journal of Nutrition* 147 (8): 1552–57.
Contreras, Alejandra V., Benjamin Cocom-Chan, Georgina Hernandez-Montes, Tobias Portillo-Bobadilla, and Osbaldo Resendis-Antonio. 2016. “Host-Microbiome Interaction and Cancer: Potential Application in Precision Medicine.” *Frontiers in Physiology*. https://doi.org/10.3389/fphys.2016.00606.
Cosentino, Francesco, Peter J. Grant, Victor Aboyans, Clifford J. Bailey, Antonio Ceriello, Victoria Delgado, Massimo Federici, et al. 2020. “2019 ESC Guidelines on Diabetes, Pre-Diabetes, and Cardiovascular Diseases Developed in Collaboration with the EASD.” *European Heart Journal* 41 (2): 255–323.
“Country Profiles.” 2016. *Trade Profiles 2016*. WTO. https://doi.org/10.30875/77418f1d-en.
Dambrova, M., G. Latkovskis, J. Kuka, I. Strele, I. Konrade, S. Grinberga, D. Hartmane, O. Pugovics, A. Erglis, and E. Liepinsh. 2016. “Diabetes Is Associated with Higher Trimethylamine N-Oxide Plasma Levels.” *Experimental and Clinical Endocrinology & Diabetes: Official Journal, German Society of Endocrinology [and] German Diabetes Association* 124 (4): 251–56.
Davis, Jesse, and Mark Goadrich. 2006. “The Relationship between Precision-Recall and ROC Curves.” *Proceedings of the 23rd International Conference on Machine Learning - ICML ’06*. https://doi.org/10.1145/1143844.1143874.
Diabetes Association, American. 2022. “Introduction: Standards of Medical Care in Diabetes—2022.” *Diabetes Care*. https://diabetesjournals.org/care/article-abstract/45/Supplement_1/S1/138921.
Diener, Christian, María de Lourdes Reyes-Escogido, Lilia M. Jimenez-Ceja, Mariana Matus, Claudia M. Gomez-Navarro, Nathaniel D. Chu, Vivian Zhong, et al. 2020. “Progressive

Shifts in the Gut Microbiome Reflect Prediabetes and Diabetes Development in a Treatment-Naive Mexican Cohort." *Frontiers in Endocrinology* 11: 602326.
Duncan, Glen E. 2006. "Exercise, Fitness, and Cardiovascular Disease Risk in Type 2 Diabetes and the Metabolic Syndrome." *Current Diabetes Reports* 6 (1): 29–35.
Esquivel-Herná, Diego Armando, Yoscelina Estrella Martínez-López, Jean Paul Sánchez-Castañeda, Daniel Neri-Rosario, Cristian Padrón-Manrique, David Girón-Villalobos, Cristian Mendoza-Ortíz, and Osbaldo Resendis-Antonio. 2022. "A Network Perspective on the Ecology of Gut Microbiota and Progression of Type 2 Diabetes: Linkages to Keystone Taxa in a Mexican Cohort." *Research Square*. https://doi.org/10.21203/rs.3.rs-1848436/v1.
Fang, Wanjun, Hongliang Xue, Xu Chen, Ke Chen, and Wenhua Ling. 2019. "Supplementation with Sodium Butyrate Modulates the Composition of the Gut Microbiota and Ameliorates High-Fat Diet-Induced Obesity in Mice." *The Journal of Nutrition* 149 (5): 747–54.
Farhangi, Mahdieh Abbasalizad, Mahdi Vajdi, and Mohammad Asghari-Jafarabadi. 2020. "Gut Microbiota-Associated Metabolite Trimethylamine N-Oxide and the Risk of Stroke: A Systematic Review and Dose-Response Meta-Analysis." *Nutrition Journal* 19 (1): 76.
Flannick, Jason, Josep M. Mercader, Christian Fuchsberger, Miriam S. Udler, Anubha Mahajan, Jennifer Wessel, Tanya M. Teslovich, et al. 2019. "Exome Sequencing of 20,791 Cases of Type 2 Diabetes and 24,440 Controls." *Nature* 570 (7759): 71–76.
Flint, Harry J., Karen P. Scott, Petra Louis, and Sylvia H. Duncan. 2012. "The Role of the Gut Microbiota in Nutrition and Health." *Nature Reviews. Gastroenterology & Hepatology* 9 (10): 577–89.
Garreta, Raul, and Guillermo Moncecchi. 2013. *Learning Scikit-Learn: Machine Learning in Python*. Packt Pub Limited.
Gong, Qiuhong, Ping Zhang, Jinping Wang, Edward W. Gregg, Yiling J. Cheng, Guangwei Li, Peter H. Bennett, and Da Qing Diabetes Prevention Outcome Study Group. 2021. "Efficacy of Lifestyle Intervention in Adults with Impaired Glucose Tolerance with and without Impaired Fasting Plasma Glucose: A Post Hoc Analysis of Da Qing Diabetes Prevention Outcome Study." *Diabetes, Obesity & Metabolism* 23 (10): 2385–94.
Grundy, Scott M. 2006. "Metabolic Syndrome: Connecting and Reconciling Cardiovascular and Diabetes Worlds." *Journal of the American College of Cardiology* 47 (6): 1093–1100.
Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585 (7825): 357–62.
Hasan, Nihal, and Hongyi Yang. 2019. "Factors Affecting the Composition of the Gut Microbiota, and Its Modulation." *PeerJ* 7 (August): e7502.
Hino, Shingo, Takayasu Mizushima, Katsunori Kaneko, Erika Kawai, Takashi Kondo, Tomomi Genda, Takahiro Yamada, Koji Hase, Naomichi Nishimura, and Tatsuya Morita. 2020. "Mucin-Derived O-Glycans Act as Endogenous Fiber and Sustain Mucosal Immune Homeostasis via Short-Chain Fatty Acid Production in Rat Cecum." *The Journal of Nutrition* 150 (10): 2656–65.
Hodgson, Kelly, Jodie Morris, Tahnee Bridson, Brenda Govan, Catherine Rush, and Natkunam Ketheesan. 2015. "Immunological Mechanisms Contributing to the Double Burden of Diabetes and Intracellular Bacterial Infections." *Immunology* 144 (2): 171–85.
Hooks, Katarzyna B., and Maureen A. O'Malley. 2017. "Dysbiosis and Its Discontents." *mBio* 8 (5). https://doi.org/10.1128/mBio.01492-17.

Hostalek, Ulrike. 2019. “Global Epidemiology of Prediabetes - Present and Future Perspectives.” *Clinical Diabetes and Endocrinology* 5 (May): 5.

Hugues, Yazmín, Rolando G. Díaz-Zavala, Trinidad Quizán-Plata, Camila Corvalán, and Michelle M. Haby. 2021. “Poor Compliance with School Food Environment Guidelines in Elementary Schools in Northwest Mexico: A Cross-Sectional Study.” *PloS One* 16 (11): e0259720.

Hussain, H. A., Y. Jin, M. Gongora, and L. J. Dunford. 2021. “Compliance of People with Diabetes in the National Diet and Nutrition Survey (2008–2016) to the National Institute for Health and Care Excellence Dietary Guidelines for Type 2 Diabetes Management.” *Proceedings of the Nutrition Society*. https://doi.org/10.1017/s0029665121003281.

“Interactions between Gut Microbiota, Host Genetics and Diet Modulate the Predisposition to Obesity and Metabolic Syndrome.” 2015. *Cell Metabolism* 22 (3): 516–30.

Iulia-Suceveanu, Andra, Sergiu Ioan Micu, Claudia Voinea, Madalina Elena Manea, Doina Catrinoiu, Laura Mazilu, Anca Pantea Stoian, Irinel Parepa, Roxana Adriana Stoica, and Adrian-Paul Suceveanu. 2019. “Metformin and Its Benefits in Improving Gut Microbiota Disturbances in Diabetes Patients.” *Metformin [Working Title]*. https://doi.org/10.5772/intechopen.88749.

Kahn, Steven E., Mark E. Cooper, and Stefano Del Prato. 2014. “Pathophysiology and Treatment of Type 2 Diabetes: Perspectives on the Past, Present, and Future.” *The Lancet* 383 (9922): 1068–83.

Karlsson, Fredrik H., Valentina Tremaroli, Intawat Nookaew, Göran Bergström, Carl Johan Behre, Björn Fagerberg, Jens Nielsen, and Fredrik Bäckhed. 2013. “Gut Metagenome in European Women with Normal, Impaired and Diabetic Glucose Control.” *Nature* 498 (7452): 99–103.

Koenig, Ronald, Clifford J. Rosen, Richard Auchus, and Allison B. Goldfine. 2019. *Williams Textbook of Endocrinology*. Elsevier.

Kostic, Aleksandar D., Ramnik J. Xavier, and Dirk Gevers. 2014. “The Microbiome in Inflammatory Bowel Disease: Current Status and the Future Ahead.” *Gastroenterology* 146 (6): 1489–99.

Kottner, Jan. 2009. “Interrater Reliability and the Kappa Statistic: A Comment on Morris et Al. (2008).” *International Journal of Nursing Studies*. https://doi.org/10.1016/j.ijnurstu.2008.04.001.

Krzewska, Aleksandra, and Iwona Ben-Skowronek. 2016. “Effect of Associated Autoimmune Diseases on Type 1 Diabetes Mellitus Incidence and Metabolic Control in Children and Adolescents.” *BioMed Research International* 2016 (July): 6219730.

Kuczynski, Justin, Jesse Stombaugh, William Anton Walters, Antonio González, J. Gregory Caporaso, and Rob Knight. 2012. “Using QIIME to Analyze 16S rRNA Gene Sequences from Microbial Communities.” *Current Protocols in Microbiology* Chapter 1 (November): Unit 1E.5.

Lakin, Steven M. 2021. “Modern Considerations for the Use of Naive Bayes in the Supervised Classification of Genetic Sequence Data.” https://search.proquest.com/openview/e0da870e6008010d0bb6b270cb11b2d2/1?pq-origsite=gscholar&cbl=18750&diss=y.

Larasati, Rahma Ayu, Dante Saksono Harbuwono, Ekowati Rahajeng, Saraswati Pradipta, Hanny Siti Nuraeni, Andi Susilowati, and Heri Wibowo. 2019. “The Role of Butyrate on Monocyte Migration and Inflammation Response in Patient with Type 2 Diabetes Mellitus.” *Biomedicines*. https://doi.org/10.3390/biomedicines7040074.

Larry Jameson, J., Anthony S. Fauci, Dennis L. Kasper, Stephen L. Hauser, Dan L. Longo,

and Joseph Loscalzo. 2018. *Harrison's Principles of Internal Medicine, Twentieth Edition (Vol.1 & Vol.2)*. McGraw-Hill Education / Medical.

Lee, Clare J., Cynthia L. Sears, and Nisa Maruthur. 2020. "Gut Microbiome and Its Role in Obesity and Insulin Resistance." *Annals of the New York Academy of Sciences* 1461 (1): 37–52.

Lee-Sarwar, Kathleen A., Jessica Lasky-Su, Rachel S. Kelly, Augusto A. Litonjua, and Scott T. Weiss. 2020. "Metabolome–Microbiome Crosstalk and Human Disease." *Metabolites*. https://doi.org/10.3390/metabo10050181.

León-Pedroza, José Israel, Luis Alonso González-Tapia, Esteban del Olmo-Gil, Diana Castellanos-Rodríguez, Galileo Escobedo, and Antonio González-Chávez. 2015. "Inflamación Sistémica de Grado Bajo Y Su Relación Con El Desarrollo de Enfermedades Metabólicas: De La Evidencia Molecular a La Aplicación Clínica." *Cirugía Y Cirujanos* 83 (6): 543–51.

Leshem, Avner, Eran Segal, and Eran Elinav. 2020. "The Gut Microbiome and Individual-Specific Responses to Diet." *mSystems* 5 (5). https://doi.org/10.1128/mSystems.00665-20.

Libbrecht, Maxwell W., and William Stafford Noble. 2015. "Machine Learning Applications in Genetics and Genomics." *Nature Reviews. Genetics* 16 (6): 321–32.

Li, Qingqing, Hui Yang, Peipei Wang, Xiaocen Liu, Kun Lv, and Mingquan Ye. 2022. "XGBoost-Based and Tumor-Immune Characterized Gene Signature for the Prediction of Metastatic Status in Breast Cancer." *Journal of Translational Medicine* 20 (1): 177.

Liu, Xuemei, Bingyong Mao, Jiayu Gu, Jiaying Wu, Shumao Cui, Gang Wang, Jianxin Zhao, Hao Zhang, and Wei Chen. 2021. "Blautia—a New Functional Genus with Potential Probiotic Properties?" *Gut Microbes* 13 (1): 1875796.

Liu, Yarong, and Min Dai. 2020. "Trimethylamine N-Oxide Generated by the Gut Microbiota Is Associated with Vascular Inflammation: New Insights into Atherosclerosis." *Mediators of Inflammation* 2020 (February): 4634172.

Lo, Chieh, and Radu Marculescu. 2019. "MetaNN: Accurate Classification of Host Phenotypes from Metagenomic Data Using Neural Networks." *BMC Bioinformatics* 20 (Suppl 12): 314.

Lundberg, Scott M., Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. "From Local Explanations to Global Understanding with Explainable AI for Trees." *Nature Machine Intelligence* 2 (1): 56–67.

Magliano, Dianna J., Edward J. Boyko, and IDF Diabetes Atlas 10th edition scientific committee. n.d. *IDF DIABETES ATLAS*. Brussels: International Diabetes Federation.

Maimon, Oded Z., and Rokach Lior. 2014. *Data Mining With Decision Trees: Theory And Applications (2nd Edition)*. World Scientific.

Mallick, Himel, Siyuan Ma, Eric A. Franzosa, Tommi Vatanen, Xochitl C. Morgan, and Curtis Huttenhower. 2017. "Experimental Design and Quantitative Analysis of Microbial Community Multiomics." *Genome Biology* 18 (1): 228.

Marchesi, Julian R., and Jacques Ravel. 2015. "The Vocabulary of Microbiome Research: A Proposal." *Microbiome*. https://doi.org/10.1186/s40168-015-0094-5.

Marcos-Zambrano, Laura Judith, Kanita Karaduzovic-Hadziabdic, Tatjana Loncar Turukalo, Piotr Przymus, Vladimir Trajkovik, Oliver Aasmets, Magali Berland, et al. 2021. "Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment." *Frontiers in Microbiology* 12 (February): 634511.

Martin-Gallausiaux, Camille, Ludovica Marinelli, Hervé M. Blottière, Pierre Larraufie, and Nicolas Lapaque. 2021. “SCFA: Mechanisms and Functional Importance in the Gut.” *Proceedings of the Nutrition Society*. https://doi.org/10.1017/s0029665120006916.

McCoubrey, Laura E., Moe Elbadawi, Mine Orlu, Simon Gaisford, and Abdul W. Basit. 2021. “Harnessing Machine Learning for Development of Microbiome Therapeutics.” *Gut Microbes* 13 (1): 1–20.

McKinney, Wes. 2010. “Data Structures for Statistical Computing in Python.” *Proceedings of the Python in Science Conference*. https://doi.org/10.25080/majora-92bf1922-00a.

Meigs, James B., Denis C. Muller, David M. Nathan, Deirdre R. Blake, Reubin Andres, and Baltimore Longitudinal Study of Aging. 2003. “The Natural History of Progression from Normal Glucose Tolerance to Type 2 Diabetes in the Baltimore Longitudinal Study of Aging.” *Diabetes* 52 (6): 1475–84.

Mejía, Lina Sofía Palacio, Jorge Leonel Wheatley Fernández, Iliana Ordoñez Hernández, Ruy López Ridaura, Hugo Lopez-Gatell Ramirez, Mauricio Hernandez Avila, Juan Eugenio Hernández Ávila, and Grupo interinstitucional para la estimacion del exceso de mortalidad. 2021. “Estimación Del Exceso de Mortalidad Por Todas Las Causas Durante La Pandemia Del Covid-19 En México.” *Salud Pública de México*. https://doi.org/10.21149/12225.

Melmed, Shlomo, Kenneth S. Polonsky, P. Reed Larsen, and Henry M. Kronenberg. 2015. *Williams Textbook of Endocrinology E-Book*. Elsevier Health Sciences.

Mohammad, Shireen, and Christoph Thiemermann. 2020. “Role of Metabolic Endotoxemia in Systemic Inflammation and Potential Interventions.” *Frontiers in Immunology* 11: 594150.

Mohan, Viswanathan, and Rajendra Pradeepa. 2017. “The Global Burden of Diabetes and Its Vascular Complications.” *Mechanisms of Vascular Defects in Diabetes Mellitus*. https://doi.org/10.1007/978-3-319-60324-7_1.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. 2018. *Foundations of Machine Learning, Second Edition*. MIT Press.

Muoio, Deborah M., and Christopher B. Newgard. 2008. “Mechanisms of disease:Molecular and Metabolic Mechanisms of Insulin Resistance and Beta-Cell Failure in Type 2 Diabetes.” *Nature Reviews. Molecular Cell Biology* 9 (3): 193–205.

Naresh, E., B. P. Vijaya Kumar, Ayesha, and Sahana P. Shankar. 2020. “Impact of Machine Learning in Bioinformatics Research.” In *Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications*, edited by K. G. Srinivasa, G. M. Siddesh, and S. R. Manisekhar, 41–62. Singapore: Springer Singapore.

Nie, Kai, Kejia Ma, Weiwei Luo, Zhaohua Shen, Zhenyu Yang, Mengwei Xiao, Ting Tong, Yuanyuan Yang, and Xiaoyan Wang. 2021. “Roseburia Intestinalis: A Beneficial Gut Organism From the Discoveries in Genus and Species.” *Frontiers in Cellular and Infection Microbiology* 11 (November): 757718.

Oh, Min, and Liqing Zhang. 2020. “DeepMicro: Deep Representation Learning for Disease Prediction Based on Microbiome Data.” *Scientific Reports* 10 (1): 6026.

Padron-Manrique, Cristian, Aarón Vázquez-Jiménez, Diego Armando Esquivel-Hernandez, Yoscelina Estrella Martinez Lopez, Daniel Neri-Rosario, Jean Paul Sánchez-Castañeda, David Giron-Villalobos, and Osbaldo Resendis-Antonio. n.d. “Mb-PHENIX: Diffusion and Supervised Uniform Manifold Approximation for Denoising Microbiota Data.” https://doi.org/10.1101/2022.06.23.497285.

Pasolli, Edoardo, Duy Tin Truong, Faizan Malik, Levi Waldron, and Nicola Segata. 2016. “Machine Learning Meta-Analysis of Large Metagenomic Datasets: Tools and Biological

Insights." *PLoS Computational Biology* 12 (7): e1004977.
Pessin, J. E., and A. R. Saltiel. 2000. "Signaling Pathways in Insulin Action: Molecular Targets of Insulin Resistance." *The Journal of Clinical Investigation* 106 (2): 165–69.
Pols, Thijs W. H., Lilia G. Noriega, Mitsunori Nomura, Johan Auwerx, and Kristina Schoonjans. 2011. "The Bile Acid Membrane Receptor TGR5: A Valuable Metabolic Target." *Digestive Diseases* 29 (1): 37–44.
Qin, Junjie, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, et al. 2012. "A Metagenome-Wide Association Study of Gut Microbiota in Type 2 Diabetes." *Nature* 490 (7418): 55–60.
Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. "The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools." *Nucleic Acids Research* 41 (Database issue): D590–96.
Rajakovich, Lauren J., Beverly Fu, Maud Bollenbach, and Emily P. Balskus. 2021. "Elucidation of an Anaerobic Pathway for Metabolism of L-Carnitine-Derived γ-Butyrobetaine to Trimethylamine in Human Gut Bacteria." *Proceedings of the National Academy of Sciences of the United States of America* 118 (32). https://doi.org/10.1073/pnas.2101498118.
Rosen, Evan D., Klaus H. Kaestner, Rama Natarajan, Mary-Elizabeth Patti, Richard Sallari, Maike Sander, and Katalin Susztak. 2018. "Epigenetics and Epigenomics: Implications for Diabetes and Obesity." *Diabetes* 67 (10): 1923–31.
Salman, Shaeke, and Xiuwen Liu. 2019. "Overfitting Mechanism and Avoidance in Deep Neural Networks." *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1901.06566.
Sanna, Serena, Natalie R. van Zuydam, Anubha Mahajan, Alexander Kurilshikov, Arnau Vich Vila, Urmo Võsa, Zlatan Mujagic, et al. 2019. "Causal Relationships among the Gut Microbiome, Short-Chain Fatty Acids and Metabolic Diseases." *Nature Genetics* 51 (4): 600–605.
Santos, Miriam Seoane, Jastin Pompeu Soares, Pedro Henrigues Abreu, Helder Araujo, and Joao Santos. 2018. "Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier]." *IEEE Computational Intelligence Magazine* 13 (4): 59–76.
Schauer, Philip R., Zubaidah Nor Hanipah, and Francesco Rubino. 2017. "Metabolic Surgery for Treating Type 2 Diabetes Mellitus: Now Supported by the World's Leading Diabetes Organizations." *Cleveland Clinic Journal of Medicine*. https://doi.org/10.3949/ccjm.84.s1.06.
Scheithauer, Torsten P. M., Elena Rampanelli, Max Nieuwdorp, Bruce A. Vallance, C. Bruce Verchere, Daniël H. van Raalte, and Hilde Herrema. 2020. "Gut Microbiota as a Trigger for Metabolic Inflammation in Obesity and Type 2 Diabetes." *Frontiers in Immunology* 11 (October): 571731.
Seki, Yoshinori, Lyda Williams, Patricia M. Vuguin, and Maureen J. Charron. 2012. "Minireview: Epigenetic Programming of Diabetes and Obesity: Animal Models." *Endocrinology* 153 (3): 1031–38.
Shamah-Levy, Teresa, Martín Romero-Martínez, Lucia Cuevas-Nasu, Ignacio Méndez Gómez-Humaran, Marco Antonio Avila-Arcos, and Juan A. Rivera-Dommarco. 2019. "The Mexican National Health and Nutrition Survey as a Basis for Public Policy Planning: Overweight and Obesity." *Nutrients* 11 (8). https://doi.org/10.3390/nu11081727.
Sharma, Sapna, and Prabhanshu Tripathi. 2019. "Gut Microbiome and Type 2 Diabetes:

Where We Are and Where to Go?" *The Journal of Nutritional Biochemistry* 63 (January): 101–8.
SIGMA Type 2 Diabetes Consortium, Amy L. Williams, Suzanne B. R. Jacobs, Hortensia Moreno-Macías, Alicia Huerta-Chagoya, Claire Churchhouse, Carla Márquez-Luna, et al. 2014. "Sequence Variants in SLC16A11 Are a Common Risk Factor for Type 2 Diabetes in Mexico." *Nature* 506 (7486): 97–101.
Sikalidis, Angelos K., and Adeline Maykish. 2020. "The Gut Microbiome and Type 2 Diabetes Mellitus: Discussing a Complex Relationship." *Biomedicines* 8 (1). https://doi.org/10.3390/biomedicines8010008.
Singh, Amrit, Casey P. Shannon, Benoît Gautier, Florian Rohart, Michaël Vacher, Scott J. Tebbutt, and Kim-Anh Lê Cao. 2019. "DIABLO: An Integrative Approach for Identifying Key Molecular Drivers from Multi-Omics Assays." *Bioinformatics* 35 (17): 3055–62.
Statnikov, Alexander, Alexander V. Alekseyenko, Zhiguo Li, Mikael Henaff, Guillermo I. Perez-Perez, Martin J. Blaser, and Constantin F. Aliferis. 2013. "Microbiomic Signatures of Psoriasis: Feasibility and Methodology Comparison." *Scientific Reports* 3: 2620.
Strandwitz, Philip, Ki Hyun Kim, Darya Terekhova, Joanne K. Liu, Anukriti Sharma, Jennifer Levering, Daniel McDonald, et al. 2019. "GABA-Modulating Bacteria of the Human Gut Microbiota." *Nature Microbiology* 4 (3): 396–403.
Stumvoll, Michael, Barry J. Goldstein, and Timon W. van Haeften. 2005. "Type 2 Diabetes: Principles of Pathogenesis and Therapy." *The Lancet* 365 (9467): 1333–46.
Thaiss, Christoph A., Maayan Levy, Inna Grosheva, Danping Zheng, Eliran Soffer, Eran Blacher, Sofia Braverman, et al. 2018. "Hyperglycemia Drives Intestinal Barrier Dysfunction and Risk for Enteric Infection." *Science* 359 (6382): 1376–83.
Thingholm, Louise B., Malte C. Rühlemann, Manja Koch, Brie Fuqua, Guido Laucke, Ruwen Boehm, Corinna Bang, et al. 2019. "Obese Individuals with and without Type 2 Diabetes Show Different Gut Microbial Functional Capacity and Composition." *Cell Host & Microbe* 26 (2): 252–64.e10.
Tönnies, Thaddäus, Wolfgang Rathmann, Annika Hoyer, Ralph Brinks, and Oliver Kuss. 2021. "Quantifying the Underestimation of Projected Global Diabetes Prevalence by the International Diabetes Federation (IDF) Diabetes Atlas." *BMJ Open Diabetes Research & Care*. https://doi.org/10.1136/bmjdrc-2021-002122.
Topçuoğlu, Begüm D., Nicholas A. Lesniak, Mack T. Ruffin 4th, Jenna Wiens, and Patrick D. Schloss. 2020. "A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems." *mBio* 11 (3). https://doi.org/10.1128/mBio.00434-20.
Torres-Leal, Francisco L., Miriam H. Fonseca-Alaniz, Marcelo M. Rogero, and Julio Tirapegui. 2010. "The Role of Inflamed Adipose Tissue in the Insulin Resistance." *Cell Biochemistry and Function* 28 (8): 623–31.
Trøseid, M., T. Ueland, J. R. Hov, A. Svardal, I. Gregersen, C. P. Dahl, S. Aakhus, et al. 2015. "Microbiota-Dependent Metabolite Trimethylamine-N-Oxide Is Associated with Disease Severity and Survival of Patients with Chronic Heart Failure." *Journal of Internal Medicine* 277 (6): 717–26.
Tsalamandris, Sotirios, Alexios S. Antonopoulos, Evangelos Oikonomou, George-Aggelos Papamikroulis, Georgia Vogiatzi, Spyridon Papaioannou, Spyros Deftereos, and Dimitris Tousoulis. 2019. "The Role of Inflammation in Diabetes: Current Concepts and Future Perspectives." *European Cardiology* 14 (1): 50–59.
Uddin, Shahadat, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. 2019. "Comparing Different Supervised Machine Learning Algorithms for Disease Prediction."

BMC Medical Informatics and Decision Making 19 (1): 281.
Unar-Munguía, Mishel, Ana Lilia Lozada-Tequeanes, Dinorah González-Castell, Manuel A. Cervantes-Armenta, and Anabelle Bonvecchio. 2021. “Breastfeeding Practices in Mexico: Results from the National Demographic Dynamic Survey 2006-2018.” *Maternal & Child Nutrition* 17 (2): e13119.
Van Rossum, Guido, and Fred L. Drake Jr. 2011. *The Python Language Reference Manual*. Network Theory.
Vidales, A. 2019. *Machine Learning Con Matlab. Técnicas de Clasificación: Análisis Clúster, Árboles de Decisión, Análisis Discriminante Y Naive Bayes*.
Vujkovic, Marijana, Jacob M. Keaton, Julie A. Lynch, Donald R. Miller, Jin Zhou, Catherine Tcheandjieu, Jennifer E. Huffman, et al. 2020. “Discovery of 318 New Risk Loci for Type 2 Diabetes and Related Vascular Outcomes among 1.4 Million Participants in a Multi-Ancestry Meta-Analysis.” *Nature Genetics* 52 (7): 680–91.
Watanabe, Suguru, Jun Kido, Mika Ogata, Kimitoshi Nakamura, and Tomoyuki Mizukami. 2019. “Hyperglycemic Hyperosmolar State in an Adolescent with Type 1 Diabetes Mellitus.” *Endocrinology, Diabetes & Metabolism Case Reports*. https://doi.org/10.1530/edm-18-0131.
Weires, M. B., B. Tausch, P. J. Haug, C. Q. Edwards, T. Wetter, and L. A. Cannon-Albright. 2007. “Familiality of Diabetes Mellitus.” *Experimental and Clinical Endocrinology & Diabetes: Official Journal, German Society of Endocrinology [and] German Diabetes Association* 115 (10): 634–40.
White, Mariel, and Simon Barquera. 2020. “Mexico Adopts Food Warning Labels, Why Now?” *Health Systems and Reform* 6 (1): e1752063.
Xie, Jiaying, Zhoujie Tong, Longfei Shen, Yuanyuan Shang, Yulin Li, Bin Lu, Weixuan Ma, Wei Zhang, and Ming Zhong. 2022. “Amylin: New Insight into Pathogenesis, Diagnosis, and Prognosis of Non-Insulin-Dependent Diabetes-Mellitus-Related Cardiomyopathy.” *Emergency and Critical Care Medicine* 2 (1): 32.
Yao, Yao, Xiaoyu Cai, Weidong Fei, Yiqing Ye, Mengdan Zhao, and Caihong Zheng. 2022. “The Role of Short-Chain Fatty Acids in Immunity, Inflammation and Metabolism.” *Critical Reviews in Food Science and Nutrition* 62 (1): 1–12.
Ying, Xue. 2019. “An Overview of Overfitting and Its Solutions.” *Journal of Physics. Conference Series* 1168 (2): 022022.
Zaccone, Giancarlo, Md Rezaul Karim, and Ahmed Menshawy. 2017. *Deep Learning with TensorFlow*.
Zhai, Shixiang, Song Qin, Lili Li, Limeng Zhu, Zhiqiang Zou, and Li Wang. 2019. “Dietary Butyrate Suppresses Inflammation through Modulating Gut Microbiota in High-Fat Diet-Fed Mice.” *FEMS Microbiology Letters* 366 (13). https://doi.org/10.1093/femsle/fnz153.
Zhao, Lijuan, Hongxiang Lou, Ying Peng, Shihong Chen, Li Fan, and Xiaobo Li. 2020. “Elevated Levels of Circulating Short-Chain Fatty Acids and Bile Acids in Type 2 Diabetes Are Linked to Gut Barrier Disruption and Disordered Gut Microbiota.” *Diabetes Research and Clinical Practice* 169 (November): 108418.
Zhou, Yi-Hui, and Paul Gallins. 2019. “A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction.” *Frontiers in Genetics* 0. https://doi.org/10.3389/fgene.2019.00579.

International Diabetes Federation.IDF Diabetes Atlas, 10th edn. Brussels, Belgium: 2021.

Available at: https://www.diabetesatlas.org