Universidad Autónoma de Querétaro
Facultad de Ingeniería
Maestría en Ciencias (Mecatrónica)

# Desarrollo de sistema de inteligencia artificial aplicado a bases de datos para diagnosticar síndrome de intestino irritable

TESIS

Que como parte de los requisitos para obtener el Grado de

Maestro en Ciencias (Mecatrónica)

Presenta:

**Israel Cinta Ramírez**

Dirigido por:

Dr. Arturo Yosimar Jaen Cuellar

Dr. Arturo Yosimar Jaen Cuellar
Presidente
Dr. Roberto Carlos Álvarez Martínez
Secretario
Dr. Miguel Trejo Hernández
Vocal
Dr. Juan Pablo Amézquita Sánchez
Suplente
Dr. Ángel Pérez Cruz
Suplente

Centro Universitario, Querétaro, Qro.
_____ (2025)
México

# Contenido

# Índice de imágenes

# Índice de tablas



# Índice de abreviaturas

IA……Inteligencia Artificial
AM……..…Aprendizaje Máquina
SII……Síndrome de Intestino Irritable
ANN-MLP……Artificial Neural Network Multi Layer Perceptron
SVM……..Support Vector Machines
RF…….....Random Forest
KNN……K Nearest Network
AI……Artificial Inteligence
ML……Machine Learning
IBS……..…Irritable Bowel Syndrome
DL……..……Deep Learning
LDA……Linear Discriminant Analysis
LASSO……....Least Absolute Shrinkage and Selection Operator
PRONACES……..……..Programas Nacionales Estratégicos
SECIHTI……..……Secretaría de Ciencias, Humanidades, Tecnología e Innovación
PCA……..……Principal Component Analysis

# RESUMEN

El desarrollo de la Inteligencia Artificial (IA) y el Aprendizaje Máquina (AM) ofrece hoy en día la posibilidad de analizar grandes volúmenes de datos y obtener información valiosa que, de otro modo, sería inalcanzable. El uso de estas metodologías se ha extendido a múltiples campos, especialmente en la medicina, donde han demostrado ser altamente fiables al analizar datos de pacientes y proporcionar diagnósticos más exactos y precisos. El presente trabajo busca reducir las limitaciones asociadas con los métodos tradicionales de diagnóstico del Síndrome de Intestino Irritable (SII), una enfermedad que afecta a una gran parte de la población mundial y cuyo diagnóstico suele requerir tiempo, recursos y no siempre resulta suficientemente preciso. El objetivo fue desarrollar una herramienta basada en IA para el diagnóstico del SII. Para ello, se utilizó una base de datos pública con información sobre las abundancias bacterianas de 39 personas (30 con SII y 9 sanas), de la cual se obtuvieron seis indicadores estadísticos de cada filo bacteriano y cinco índices de diversidad por paciente. Con estos datos se entrenaron cuatro métodos de AM: Redes Neuronales Artificiales Multicapa (ANN-MLP), Máquinas de Soporte Vectorial (SVM), Bosque Aleatorio (RF) y K Vecinos más Cercanos (KNN). Los modelos se evaluaron mediante matrices de confusión, obteniendo métricas de exactitud, precisión, sensibilidad y puntaje F1. Aunque el desempeño de los métodos varió, todos mostraron resultados satisfactorios; por ello, la herramienta desarrollada permite trabajar con los cuatro algoritmos, ofreciendo adaptabilidad si el usuario desea emplear diferentes bases de datos. En general, todos los métodos alcanzaron una exactitud del 92%; sin embargo, los mejores resultados se obtuvieron con las ANN-MLP y las SVM, que lograron una precisión del 100%, superando al método de diagnóstico más utilizado actualmente, el criterio de Roma IV, cuya precisión es del 82.4%.

**Palabras clave**: Inteligencia Artificial, análisis de datos, aprendizaje máquina, diagnóstico, síndrome de intestino irritable.

# ABSTRACT

The development of Artificial Intelligence (AI) and Machine Learning (ML) now offers the ability to analyze large volumes of data and extract valuable information that would otherwise be unattainable. The application of these methodologies has expanded across multiple fields, particularly in medicine, where they have proven to be highly reliable in analyzing patient data and providing more accurate and precise diagnoses. The present work aims to reduce the limitations associated with traditional diagnostic methods for Irritable Bowel Syndrome (IBS), a condition that affects a large portion of the global population and whose diagnosis often requires significant time and resources while lacking sufficient accuracy. The objective was to develop an AI-based tool for IBS diagnosis. For this purpose, a public database containing bacterial abundance data from 39 individuals (30 with IBS and 9 healthy) was used, from which six statistical indicators for each bacterial phylum and five diversity indices per patient were obtained. Using these data, four ML algorithms were trained: Artificial Neural Networks – Multilayer Perceptron (ANN-MLP), Support Vector Machines (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN). The models were evaluated using confusion matrices, obtaining metrics such as accuracy, precision, sensitivity, and F1-score. Although the performance of the methods varied, all produced satisfactory results; therefore, the developed tool allows the use of all four algorithms, offering flexibility when working with different datasets. Overall, all models achieved an accuracy of 92%; however, the best results were obtained with ANN-MLP and SVM, which reached a precision of 100%, surpassing the most commonly used diagnostic approach for IBS, the Rome IV criteria, which reports a precision of 82.4%.



**Key words**: Artificial Intelligence, data analysis, machine learning, diagnosis, irritable bowel syndrome.

# AGRADECIMIENTOS

A la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), por el otorgamiento de la beca con número de CVU 1316166, la cual fue fundamental para la realización y conclusión del presente trabajo de investigación. Asimismo, expreso mi profundo agradecimiento por el apoyo brindado, que permitió cumplir con los objetivos del programa de posgrado de la Maestría en Ciencias (Mecatrónica).

A la Universidad Autónoma de Querétaro, por haberme formado desde los 18 años y por brindarme las herramientas necesarias para incorporarme al mercado laboral.

De igual manera, manifiesto mi sincero agradecimiento a mi asesor, sinodales y profesores, quienes con su orientación, conocimientos y compromiso contribuyeron de manera significativa al desarrollo de este trabajo y a mi formación académica. Su guía y consejo fueron pilares esenciales en la consecución de estos logros.

Se agradece a EMBL-EBI por compartir la base de datos con la que se trabajó en el presente trabajo.

# DEDICATORIA

Para alguien de pocas palabras, esta es una gran oportunidad para expresar aquello que normalmente solo se piensa. Así que aprovecho para agradecer:

A mi madre, verdaderamente la mejor madre del mundo. Un sol que ilumina a todos los que la conocen y quien me enseñó todo lo que está bien y lo que está mal.

A mi padre, quien me mostró lo que es el sacrificio y lo que significa ser un buen hombre para la sociedad.

A mi hermana, con quien he compartido enojos, llantos y risas desde la infancia, y que siempre ha llenado nuestros días de vida y alegría.

A mi vieja amiga Madame Nekoi, la persona en quien más confío. Espero que nada nos separe hasta morir de viejos.

A mis amigos de la carrera, porque tener un compañero a tu lado hace que el infierno sea más llevadero, en especial a cierta caballera de Izalith, cuya voluntad admiro profundamente.

A mis compañeros del posgrado, todos tan únicos como locos y siempre confiables; en especial a ese amigo de la familia Ursidae, porque conocer a alguien tan transparente te devuelve la fe en la gente.

# CAPÍTULO 1: INTRODUCCIÓN

El desarrollo de técnicas de IA ha permitido el análisis de grandes cantidades de datos para la toma de decisiones, razón por la cual se ha empezado a utilizar en distintos campos, incluso en el campo de la medicina. En el ámbito médico, la detección de algunas enfermedades presenta múltiples desafíos, ya que depende de factores como la experiencia del médico, la correcta identificación de indicadores biológicos, y la similitud con otras patologías (Norman et al. 2017). Estos desafíos contribuyen a que muchas enfermedades no sean diagnosticadas adecuadamente y a que el proceso se vuelva largo y costoso. Este es el caso del SII, una afección que impacta a miles de personas en todo el mundo y que carece de marcadores biológicos claros para un buen diagnóstico (Bassotti 2022). Por ello, en este trabajo se propuso una metodología basada en la aplicación de cuatro técnicas de Aprendizaje de Máquina (Machine Learning, ML) en bases de datos libres ya existentes de abundancias bacterianas del intestino, de donde se obtuvieron cinco índices de diversidad biológica e indicadores estadísticos de cada filo de bacterias para diagnosticar el SII. Se calificó el mejor de los cuatro métodos a través de métricas de desempeño, las cuales fueron precisión, exactitud, sensibilidad y F1. Se desarrolló un código para un programa con el fin de que cualquier profesional de la salud o investigador interesado pudiera introducir los datos de abundancias de su paciente o usuario en un archivo con formato de valores separados por comas (.csv) y obtener un diagnóstico y las razones de este diagnóstico. Al hacerlo, se buscó no solo evitar las complicaciones asociadas con las imprecisiones, incertidumbres y limitaciones de los métodos de diagnóstico tradicionales, sino también proporcionar información valiosa sobre las variables que influyeron en el diagnóstico. Así entonces, el método propuesto fue alimentado con información proporcionada por indicadores biológicos extraídos de bases de datos de acceso libre sobre la diversidad bacteriana

de los usuarios, la cual fue procesada posteriormente mediante las técnicas de ML para identificar patrones que sirvieran como discriminadores entre usuarios sanos y aquellos con SII.

## 1.1 ANTECEDENTES

Hoy en día, el desarrollo de metodologías modernas que permitan diagnosticar procesos y sistemas en distintos campos de la investigación resulta de gran interés, debido a la necesidad de obtener resultados precisos y confiables bajo esquemas robustos que resuelvan problemáticas con características complejas. En este sentido, se han desarrollado diversas técnicas que permiten el monitoreo y análisis de sistemas a partir de datos obtenidos de diversas fuentes, y que posteriormente son almacenados para llevar a cabo su procesamiento, permitiendo generar modelos, predicciones y diagnósticos. Dentro de las metodologías de gestión de datos se encuentra una rama conocida como ML, que integra el procesamiento de datos, la extracción de indicadores y la aplicación de técnicas de clasificación para producir un diagnóstico de salida. En otras palabras, estas estrategias permiten llevar a cabo procesos de discriminación de eventos, así como la detección de cambios o anomalías en el sistema. Cabe destacar que estas metodologías de diagnóstico pueden aplicarse en diversos campos de la ciencia, tales como en ingeniería, química, medicina, entre otros; en el caso del presente trabajo, se hace referencia a su aplicación en el área de la biología, específicamente en el ámbito de la salud.

El diagnóstico médico constituye la principal herramienta para tratar a un paciente que presenta alguna dolencia o enfermedad, pues se fundamenta en hechos y evidencia empírica con el propósito de acercarse lo más posible a la verdad. Según Cruz et al. (2012), los principios del diagnóstico incluyen: realizar un resumen objetivo del caso, organizar la

información, jerarquizar los síntomas y signos de acuerdo con su sensibilidad, especificidad y valor predictivo, agrupar los síntomas y signos identificados, distinguir entre síndromes “duros” (los más relevantes) y “blandos”, evitar la hipertrofia diagnóstica y avanzar del síntoma y signo hacia el síndrome. En esta misma línea, Díaz et al. (2006) señalan que el médico recurre a diversas herramientas para establecer un diagnóstico, entre ellas la observación de la evidencia visible y la indagación sobre los síntomas del paciente. Asimismo, destacan la importancia de considerar factores como la edad, el sexo, la ascendencia y el historial médico con el fin de diferenciar el síndrome o síndromes actuales de otros posibles. Dichos autores también señalan que entre el 50% y el 75% de los diagnósticos se obtienen mediante el interrogatorio complementado con los exámenes clínicos que el médico estime necesarios para confirmar o replantear la hipótesis diagnóstica. En conclusión, subrayan que la combinación de diferentes métodos fortalece las valoraciones médicas, y que, aunque los avances tecnológicos contribuyen significativamente al diagnóstico estos no sustituyen el conocimiento experto del profesional de la salud. Por su parte, Lorenzano (2006) enfatiza que el campo de la medicina —y particularmente el diagnóstico médico— se apoya en hechos verificables, pero también en teorías y conjeturas influidas por la experiencia, formación y contexto del médico y del paciente. Entre sus conclusiones, destaca que los avances científicos y tecnológicos permiten reducir sesgos y diagnosticar enfermedades incluso antes de la aparición de síntomas graves; sin embargo, advierte que estos avances no reemplazarán nunca la interacción médico–paciente. En este marco de referencia, el presente trabajo propone una metodología para el desarrollo de una herramienta tecnológica en software, basada en el procesamiento de datos, que sirva como apoyo y complemento en el diagnóstico del SII. Al igual que los estudios complementarios (p. ej., análisis de sangre o ecografías), esta herramienta busca aportar mayor confiabilidad a los resultados, dado que métodos tradicionales como el interrogatorio, la observación de evidencia

visible o la revisión del historial médico, aunque útiles, pueden no ser suficientes para obtener un diagnóstico plenamente confiable.

Por su parte, la IA ha sido utilizada en diversos trabajos para facilitar el análisis de datos de pacientes, permitiendo obtener diagnósticos de manera más precisa, o procesar grandes volúmenes de información en un corto periodo de tiempo. Según Álvarez et al. (2021), en el área de la salud existen múltiples enfermedades sobre las cuales aún se conoce muy poco, ya que los estudios convencionales no siempre ofrecen la información necesaria. Debido a lo anterior, estudios recientes enfocados en el diagnóstico recurren a enfoques alternativos, como el análisis de la microbiota y su relación con las enfermedades que padece un individuo. Por ejemplo, Golondrino (2020) empleó ML para la clasificación taxonómica de bacterias, logrando una solución más rápida y eficiente para analizar secuencias de ADN. En dicho estudio, la IA permitió automatizar un proceso complejo, largo y propenso a errores humanos. De manera más aplicada, González y Pinzón (2023) recopilaron investigaciones en las que se aplicaron modelos de IA para prevenir la enfermedad periodontal. Los algoritmos más empleados de acuerdo con su investigación fueron las Máquinas de Soporte Vectorial (*Support Vector Machine* o SVM) y el Aprendizaje Profundo (*Deep Learning* o DL), aunque señalan que la elección del método depende de la aplicación específica. Por su parte, Huang et al. (2019) utilizaron cinco algoritmos de ML para analizar proteínas presentes en líquido gingival con el fin de diagnosticar la enfermedad periodontal, las técnicas utilizadas fueron: SVM, Bosque Aleatorio (*Random Forest* o RF), k Vecinos más Cercanos (*k-Nearest Neighbours* o KNN), Análisis Discriminante Lineal (*Linear Discriminant Analysis* o LDA) y Árboles de Decisión. De los cinco algoritmos, la técnica LDA mostró el mejor desempeño en este caso. De manera similar, Zia (2021) aplicó algoritmos como KNN, Árboles de Decisión, RF, Regresión Logística, SVM y Redes Neuronales Artificiales (*Artificial Neural Networks* o

ANN) para diagnosticar leucemia a partir de imágenes de frotis, alcanzando una precisión de hasta el 100% con RF y SVM. Otro estudio relevante es el de Aguirre et al. (2021), en el que un análisis de imágenes 2D para la detección automática de hemorragias y microhemorragias alcanzó una sensibilidad del 93.16%. Además, reportan cuáles modelos de IA entrenados para detectar nódulos pulmonares lograron una sensibilidad del 99.1% y una especificidad del 99.2%. Los resultados de los trabajos analizados hasta ahora reflejan la capacidad de la IA para diagnosticar enfermedades con alta precisión, aunque la mayoría de los enfoques se limitan a determinar la presencia o ausencia de la enfermedad, sin proporcionar información sobre el grado de severidad ni un diagnóstico preventivo. Así entonces, en un enfoque diferente, Víctor (2022) abordó el análisis de la microbiota como un sistema altamente complejo, que puede ser utilizado como herramienta de monitoreo y diagnóstico, debido a la interacción de millones de microorganismos. Para su análisis recurrió al uso de las Redes Neuronales Artificiales Multicapa (MANN), con el fin de identificar microorganismos compartidos por dos individuos dentro de una misma cadena trófica. En este caso, solo se aplicó MANN sin comparación con otros algoritmos. Por otro lado, más enfocado en enfermedades gastrointestinales, Zand et al. (2022) propusieron un modelo con 108 variables —incluyendo hospitalizaciones, comorbilidades y pruebas de laboratorio— para entrenar una IA destinada a identificar enfermedades inflamatorias intestinales. Para lograr su objetivo, se aplicaron la regresión Ridge, regresión LASSO (*Least Absolute Shrinkage and Selection Operator*), SVM, RF y ANN. Además, observaron que los mejores resultados se obtuvieron con RF y regresión LASSO, alcanzando precisiones entre 0.70 y 0.92. En resumen, dicho trabajo, aunque automatiza el proceso de diagnóstico, se fundamenta en la evaluación de hábitos y antecedentes médicos sin recurrir a variables cuantificables. Por su lado, en la presente propuesta se busca utilizar bases de datos ya existentes obtenidas, en su momento, mediante una sola toma de muestra por persona,

lo que evita la necesidad de realizar pruebas de laboratorio adicionales o un análisis exhaustivo de los antecedentes médicos, por lo que sería un proceso más rápido y sencillo. De manera semejante, Patón (2021) empleó IA para distinguir entre pacientes sanos y aquellos con cáncer de colon, mediante el análisis de la microbiota. Su investigación identificó variaciones en la abundancia de 40 especies bacterianas (20 en aumento y 20 en disminución) en pacientes con cáncer. Aunque este trabajo permitió establecer un perfil de disbiosis asociado a la enfermedad, no se desarrolló ni se publicó un software que pudiera emplearse como herramienta de diagnóstico clínico. Un trabajo enfocado en el uso de IA para la detección del SII es el de Fukui et al. (2020), quienes emplearon el método de RF para diferenciar a personas con SII de individuos sanos, utilizando una base de datos de 111 pacientes. Obtuvieron resultados con una sensibilidad superior al 80% y una precisión superior al 90%. En el presente trabajo se buscó utilizar el método de RF junto con ANN-MLP, SVM y KNN, con el fin de comparar su desempeño. Además, se propuso entrenar estos métodos con una base de datos sintética para posteriormente clasificar una base de datos real y, a partir de los resultados, desarrollar una herramienta de diagnóstico del SII que pueda ser utilizada por profesionales de la salud. De manera similar, Su et al. (2022) aplicaron los métodos de RF, ANN-MLP, SVM y KNN para distinguir entre personas sanas y pacientes con adenomas colorrectales, enfermedad de Crohn, cáncer colorrectal, enfermedades cardiovasculares, síndrome post-COVID-19, colitis ulcerosa y SII, utilizando una base de datos con 1038 casos. Obtuvieron resultados de exactitud del 98%, sensibilidad del 94% y precisión del 98% para el SII. No obstante, su estudio no se centró exclusivamente en esta enfermedad, ni empleó indicadores estadísticos para la condensación de información, ni ofreció una herramienta práctica que facilite la aplicación de estos métodos en diagnósticos futuros.

Particularizando en el tema, el SII es un trastorno que afecta a una gran parte de la población mundial. Según Domingo (2021), el SII es uno de los trastornos intestinales más comunes y se caracteriza por dolor abdominal, meteorismo con distensión abdominal y alteraciones en las evacuaciones intestinales, con predominio de diarrea, estreñimiento o alternancia de ambos. Por su parte, Shaikh et al. (2023) reportan datos relevantes sobre su prevalencia global: entre el 10% y el 25% en Estados Unidos, entre el 17% y el 21% en América del Sur, del 7% al 9% en Asia del Sur y aproximadamente un 5.6% en Medio Oriente y África. Además, el SII puede presentarse en todas las edades, aunque es más frecuente en personas de entre 32 y 38 años, disminuyendo su prevalencia en mayores de 50 años. En este mismo contexto, Otero y Gómez (2005) señalan que el SII constituye la segunda causa de ausentismo laboral después de la gripe, tan solo en Estados Unidos provoca entre 2.4 y 3.5 millones de consultas médicas anuales, alrededor de 2.2 millones de prescripciones y un gasto estimado de 8 mil millones de dólares en costos directos. Así mismo, estos autores subrayan que el diagnóstico del SII se realiza por exclusión, dado que no existen biomarcadores ni pruebas específicas, por lo que se descartan otras enfermedades tras múltiples exámenes. De manera complementaria, Remes et al. (2010) destacan que la cronicidad de los síntomas y la prevalencia del SII generan un elevado gasto económico, ya que en muchos casos únicamente se tratan los síntomas. Estos autores también enfatizan que el diagnóstico se basa en la sintomatología, debido a la ausencia de indicadores biológicos. No obstante, se han desarrollado herramientas diagnósticas como los Criterios de Roma, los cuales consisten en una serie de preguntas sobre los síntomas del paciente durante un período de al menos tres meses. Por ejemplo, Alvarado et al. (2015) reportaron que el Criterio de Roma III presenta una sensibilidad del 68% y una especificidad del 79%. En cambio, Black et al. (2020) hallaron que el Criterio de Roma IV alcanza una sensibilidad del 82.4% y una especificidad del 82.9%. El presente trabajo

busca mejorar estas métricas diagnósticas y reducir el tiempo de diagnóstico. A pesar de la alta prevalencia del SII, diversos autores coinciden en que aún no existe suficiente información sobre este síndrome. En este sentido, Ghaffari et al. (2022) señalan que el SII se asocia con cambios significativos en la microbiota intestinal; sin embargo, la información disponible aún no es suficiente para establecer un tratamiento adecuado. De manera similar, Chong et al. (2019) recopilan evidencia sobre alteraciones en diferentes familias y especies bacterianas en pacientes con SII, identificando qué variaciones están asociadas a cada subtipo del síndrome. Siguiendo la misma línea, Icaza (2014) describe que los pacientes con SII presentan un aumento de hasta dos veces en la relación Firmicutes/Bacteroidetes, además de una disminución en las bacterias *Lactobacillus* y *Bifidobacterium*. Asimismo, se asocia la presencia de una nueva especie, *Ruminococcus*, con la aparición del síndrome. Tanto el trabajo de Chong et al. (2019) como el de Icaza (2014) evidencian que el estudio de las abundancias bacterianas puede aportar a un mejor perfil de datos para lograr el diagnóstico del SII. Sin embargo, actualmente debido a la gran cantidad de datos involucrados, el análisis realizado de forma manual se limita a un número reducido de especies bacterianas.

Por lo tanto, el análisis de la composición de la microbiota intestinal se considera un método útil para obtener información relevante sobre la salud de una persona, aunque llevarlo a cabo puede ser complejo. Sobre esto, Álvarez et al. (2021) destacan la importancia de la microbiota intestinal debido a sus funciones en el desarrollo somático, la nutrición y la inmunidad, entre otros procesos. Asimismo, señalan que la pérdida de riqueza en determinadas especies bacterianas se relaciona con diversas enfermedades crónicas. En el presente trabajo se retoma esta idea para analizar la pérdida de riqueza microbiana en pacientes con SII y evaluar si este parámetro pudiera ser relevante en el diagnóstico. En la misma línea, Sakamoto et al.

(2022) introducen el concepto de disbiosis, entendido como el cambio en el estado estacionario de la microbiota intestinal que altera el equilibrio microbiano y se asocia con distintas patologías. Este fenómeno suele manifestarse en la disminución de bacterias comensales y el aumento de grupos potencialmente patógenos. De forma complementaria, Álvarez et al. (2021) distinguen entre eubiosis —equilibrio microbiano caracterizado por la presencia de bacterias benéficas— y disbiosis, definida como el aumento de bacterias patógenas y la disminución de bacterias protectoras. Además, relacionan la disbiosis intestinal con enfermedades como enterocolitis necrosante, sepsis neonatal, diabetes mellitus y la enfermedad inflamatoria intestinal. No obstante, aclaran que en muchos casos la evidencia no es suficiente para determinar si la disbiosis precede a la enfermedad o viceversa. Por su parte, Moreno (2022) coincide en que no siempre existe una relación directa entre disbiosis y enfermedad, aunque estudios recientes perfilan asociaciones entre ciertos patrones de disbiosis y enfermedades gastrointestinales. En este sentido, tanto Sakamoto et al. (2022) como Moreno (2022) plantean la hipótesis de que la disbiosis, ya sea como causa o consecuencia, podría constituir un indicador biológico útil para el diagnóstico de este tipo de patologías. Mientras tanto, el trabajo de Arce (2020) analizó las causas de la disbiosis y su vínculo con distintas enfermedades, para después proponer diversas terapias dirigidas a restaurar el equilibrio microbiano. La autora concluye que la microbiota intestinal y su relación con la salud del huésped sigue siendo un campo complejo y poco explorado. En concordancia, Machado et al. (2023) señalan que la microbiota regula procesos esenciales como la digestión de los alimentos, la absorción de nutrientes, la síntesis de vitaminas y ácidos biliares, la prevención de organismos patógenos, así como la expresión inmunológica y genética del huésped. Estos autores sugieren que el desbalance bacteriano podría explicar la predisposición a la obesidad en algunas personas y plantean que la manipulación de la microbiota podría constituir una alternativa terapéutica

para la obesidad y otros trastornos metabólicos. En esta misma línea, Fontané et al. (2018) recopilan estudios en los que la microbiota intestinal se modifica mediante la administración de probióticos con el objetivo de tratar la obesidad en humanos y animales, así como otros trastornos asociados. Por otro lado, Chan et al. (2013) discuten las dificultades para modificar la microbiota a través de la dieta, señalando que este campo aún se encuentra en desarrollo. No obstante, proponen un enfoque innovador denominado bacterioterapia, que consiste en la administración de cepas específicas de probióticos para prevenir, tratar o incluso curar enfermedades como el SII, el estreñimiento, el síndrome metabólico y la obesidad. En este contexto, la relación entre disbiosis y SII adquiere relevancia diagnóstica. Aunque no está del todo claro si la disbiosis es causa o consecuencia del síndrome, su identificación podría contribuir a perfilar biomarcadores y, a largo plazo, servir como base para el diseño de tratamientos. Finalmente, algunos estudios han empleado índices de diversidad microbiana como herramienta para caracterizar la microbiota. Tal es el caso de Stoppani (2013), quien calculó el índice de diversidad de Shannon, el índice de homogeneidad de Pielou y el índice de riqueza de Margalef para evaluar los cambios microbianos en lechones alimentados con probióticos. De manera similar, Velasco et al. (2018) investigaron la microbiota de conejos en relación con sus taxones, empleando índices de diversidad alfa como el número total de OTUs observados, el índice de Shannon y el inverso de Simpson. En el presente trabajo se propone integrar estos índices al análisis basado en IA para determinar su relevancia en el diagnóstico del SII.

El presente trabajo tiene como objetivo analizar una o varias bases de datos de acceso libre que contienen información sobre las abundancias bacterianas de la microbiota de individuos con diversas enfermedades, así como datos relacionados con su alimentación y hábitos de vida. A partir de esta información, se calcularán índices de diversidad microbiana que servirán

como variables adicionales para el análisis. Posteriormente, estos datos e índices se emplearán para entrenar un sistema basado en IA utilizando distintos algoritmos de aprendizaje, tales como Árboles de Decisión (RT), SVM, KNN y ANN. El propósito es que el sistema aprenda a diferenciar entre individuos sanos y personas con SII, e incluso pueda identificar patrones que permitan diagnosticar a quienes estén en riesgo de desarrollar el SII. Además, se pretende extraer los pesos que asigna el sistema de IA durante el diagnóstico, ya que estos reflejan las características del paciente y la disbiosis específica asociada al SII, lo que podría constituir un indicador biológico útil para el desarrollo de futuros tratamientos. Una vez implementados los distintos algoritmos, se evalúa su desempeño mediante métricas específicas para seleccionar el método más eficiente. Posteriormente, se desarrolla un programa de software de uso abierto basado en el algoritmo seleccionado. Esta herramienta tecnológica permite a los profesionales de la salud realizar diagnósticos del SII de manera más rápida y precisa, superando las limitaciones de los métodos tradicionales, como la exclusión de enfermedades mediante múltiples estudios o el análisis exhaustivo de la historia clínica del paciente.

## 1.2 JUSTIFICACIÓN

La justificación del presente trabajo genera impactos desde diversos puntos de vista, que a continuación se explican en detalle:

### 1.2.1 Impacto Científico

- Se desarrolla una metodología basada en técnicas de IA y ML que permita automatizar el procesamiento de bancos de datos relacionados con la composición del microbiota intestinal de usuarios para el diagnóstico del SII de manera confiable.

- Se busca extraer indicadores cuantitativos de los bancos de datos que puedan aportar información valiosa para discriminar la existencia del SII.
- El encontrar la disbiosis específica del SII podría servir como herramienta discriminatoria efectiva para el monitoreo y diagnóstico del SII.
- El aplicar técnicas de IA permite explorar y reforzar el diagnóstico del SII a través de la implementación de estrategias de ML recurriendo a la extracción de indicadores e implementando clasificadores para el diagnóstico del síndrome.

### 1.2.2 Impacto Tecnológico

- Se desarrolla un programa en software para el diagnóstico del SII a partir del procesamiento de datos con características tales como modularidad, arquitectura abierta, y licencia libre, que permita un incremento posterior de sus capacidades y portabilidad de aplicación.
- Se implementan técnicas de IA y ML, así como la extracción de indicadores que servirán para la discriminación del SII y su diagnóstico.

### 1.2.3 Impacto Económico

- Un sistema de software que permita el diagnóstico del SII basado únicamente en procesamiento de datos impacta de forma positiva en la reducción de los gastos derivados de los análisis complejos que se desarrollan de forma convencional para diagnosticar este problema de salud, es decir, se evita que se gaste en recursos innecesarios.
- El tener un sistema de diagnóstico más preciso ayuda a que no haya falsos positivos de otras enfermedades evitando gastos en tratamientos innecesarios.

### 1.2.4 Impacto Social

- El diagnosticar de manera más efectiva el SII ayuda a que se puedan tomar las medidas precisas para mejorar el estado de salud de una persona.
- El realizar diagnóstico que puede ser incluso preventivo puede ayudar a la persona que sufre SII, o que está a punto de contraerlo, a tomar las medidas necesarias en favor de su salud.
- El trabajo presente busca atender problemáticas desde el primer nivel de atención, que se conforman por la atención, prevención y promoción de la salud, enfocándose en la prevención y la atención de un síndrome que afecta hasta un 21% de la población latinoamericana, lo cual se alinea con los Programas Nacionales Estratégicos (PRONACES) establecidos por la Secretaría de Ciencias, Humanidades, Tecnología e Innovación (SECIHTI), específicamente en el rubro de la salud al proporcionar un método de diagnóstico más efectivo que el actual.

## 1.3 DESCRIPCIÓN DEL PROBLEMA

Las metodologías de diagnóstico convencionales generalmente aplican estudios de encuesta, consideran historial clínico, o ejecutan pruebas experimentales de laboratorio para descartar otras enfermedades. Sin embargo, la exploración de metodologías alternativas no-invasivas basadas en gestión y procesamiento de datos, y que además integren el poder de las técnicas de IA y ML no han sido completamente tomadas en cuenta. A continuación, se describe la problemática existente del análisis de antecedentes desde diversos puntos de vista.

### 1.3.1 Científico

- La metodología para diagnosticar el SII sigue siendo realizada de manera poco precisa, por medio de descarte de otras enfermedades o, en

el caso donde se obtiene un resultado más preciso, relacionándolo con el historial médico del paciente y de su familia.

- El que no haya indicadores biológicos conocidos del SII provoca que no haya un diagnóstico preciso ni tampoco un tratamiento eficiente, aunque sea uno de los síndromes que más afecta a la población.
- Para realizar un análisis profundo del SII es necesario analizar una gran cantidad de datos, algo que podría ser demasiado difícil para un investigador o incluso un grupo de investigadores.
- La integración de técnicas de IA y ML han demostrado resultados congruentes y efectivos en distintas áreas del conocimiento, por lo tanto, explorar su efectividad en el área de la salud para diagnosticar el SII es un área de oportunidad.
- Ajustar estas técnicas de IA y ML para diagnosticar el SII implica un reto al no tener antecedentes directos, ya que se deben definir indicadores cuantitativos que permitan a estas técnicas converger a un resultado valido.

### 1.3.2 Tecnológico

- Para encontrar un indicador biológico en el que basar un diagnóstico, o incluso un tratamiento, se puede realizar un análisis de la microbiota, pero esta tarea puede ser demasiado complicada como para que un investigador lo haga de forma manual, debido a la cantidad de variables y datos que hay que analizar, por lo que es necesario aplicar herramientas de IA.
- Definir una metodología o esquema que ejecute el procesamiento de la información de forma automática es un reto, ya que se deben integrar conceptos del área de la salud con técnicas de ML y su implementación mediante programación.
- Hoy en día no existen métodos o herramientas tecnológicas para el diagnóstico del SII, pero el uso de IA podría ser utilizado para esto.

### 1.3.3 Económico

- Hay un gasto constante en medicación de los síntomas del SII debido a su permanencia.
- El SII es la segunda causa de ausencia laboral, después de la gripe, por lo que causa que se produzcan pérdidas económicas indirectas considerables a las empresas.
- El diagnóstico del SII puede ser un gasto significativo ya que es necesario descartar otras enfermedades, lo cual se realiza mediante estudios complementarios.

### 1.3.4 Social

- El aplicar el uso de ML e IA en otros campos, en este caso de la medicina, puede sentar una base para que la sociedad se acostumbre al uso de estas y las aplique en otros ámbitos de la ciencia y la industria.
- La metodología para diagnosticar el SII sigue siendo poco precisa, ya que para su diagnóstico es necesario eliminar otras opciones mediante estudios clínicos o realizar un análisis muy amplio de la historia clínica del paciente, sus hábitos e incluso de su familia, lo cual ocasiona que haya mucha gente que sufra las consecuencias de este síndrome sin saber que lo tiene.
- El SII afecta a una gran cantidad de personas en México y el resto del mundo, un 4.1% de la población mundial y hasta un 21% de la población latinoamericana y muchas personas se ven afectadas de manera crónica, por lo que sufren de los malestares que provoca el SII como el dolor abdominal, el estreñimiento, la distención abdominal y diarrea, por lo que su calidad de vida se ve reducida.
- El SII afecta de manera crónica a una gran cantidad de personas mientras que a otros le afecta de manera intermitente.

## 1.4 HIPOTESIS

Es posible extraer información discriminante valiosa a partir de tablas de datos de pacientes, expresada en forma de indicadores estadísticos o biológicos, que, mediante un procesamiento posterior con técnicas de IA y ML, permitan diagnosticar el SII con una precisión superior a la obtenida con las metodologías actualmente más utilizadas, los criterios de Roma III y IV, con los cuales se obtienen valores de precisión máximos del 82.4%.

## 1.5 OBJETIVOS

### 1.5.1 Objetivo general

Desarrollar una metodología para el diagnóstico del SII, implementada como una herramienta en software que distinga entre un paciente saludable y uno con el síndrome, basada en el procesamiento y la gestión de bancos de datos de la composición de su microbiota intestinal, el cómputo y extracción de indicadores biológicos, y en la aplicación de algoritmos de IA y ML.

### 1.5.2 Objetivos específicos

- Seleccionar una base de datos libre existente en línea que haya sido creada por universidades y/o grupos de investigación privados. De esta manera se tiene la información necesaria para entrenar el sistema de clasificación.

- Obtener cinco índices de diversidad a través de las tablas de abundancias de bacterias para reducir el número de variables y facilitar el análisis, obteniendo a su vez métricas que proporcionen información discriminante del síndrome analizado.

- Encontrar el método de los cuatro propuestos que mejor resultados otorgue a la hora de realizar el análisis para diagnosticar el SII (KNN, SVM, RF, ANN) utilizando matrices de confusión para hacer las comparaciones de desempeño “accuracy”, “recall”, “precisión” y “F1”. El método con mejores métricas se recomienda para que sea utilizado por el usuario en la herramienta.

- Encontrar cuales son los filos de bacterias que tienen mayor peso en el análisis que realiza software para diagnosticar el SII.
- Establecer si existe una relación entre los índices de diversidad y el grado de severidad del SII.

- Diseñar un programa para el diagnóstico del SII, usando el método de IA con mejores resultados, que esté listo para ser usado por un profesional de la salud usando los datos de abundancias de bacterias del paciente en formato .csv.

# CAPITULO 2: FUNDAMENTACIÓN TEÓRICA

En esta sección se describen los temas necesarios para el desarrollo del proyecto. A continuación, se mencionan los temas principales:

- **Inteligencia Artificial.** Es una descripción de las bases del programa y de los métodos que se utilizarán para el desarrollo de este. Después, se describen los métodos que se utilizarán, entre los que están:

  - RF
  - SVM
  - KNN
  - ANN

- **Métodos para visualización.** Se muestran los métodos que se utilizan para la visualización de los datos al reducir la dimensionalidad de los datos utilizados. Los dos métodos que se utilizan son:

  - Análisis Discriminante Lineal (LDA)
  - Análisis de Componentes Principales (PCA)

- **Análisis de varianza.** Es el método que se utiliza para obtener las mejores características, con las cuales se entrenaran los algoritmos de IA.

- **Clasificación de los algoritmos.** Describen la metodología y los factores para evaluar los métodos de IA.

- **Bases biológicas.** Se describe la naturaleza de las variables biológicas con las que se está trabajando, así como el fenómeno biológico que se está analizando mediante la metodología de este trabajo.

- **Índices de diversidad.** Son las representaciones matemáticas que se utilizan para medir las variables biológicas de la base de datos que se utiliza.

## 2.1 Inteligencia artificial (IA) y técnicas de aprendizaje de máquina (AM)

Se habla de IA cuando un programa de computadora tiene la capacidad de "aprender", es decir que el programa cambia automáticamente, sin la necesidad de ninguna reprogramación o modificación extra, y puede modificar sus características para realizar su propósito de mejor manera. Cuando se habla de IA se suele hablar también del AM, o *ML* por sus siglas en inglés, que es un método basado en la supervisión de la aplicación de un algoritmo de aprendizaje con una arquitectura específica para la clasificación de datos. Entre los algoritmos de aprendizaje están los RF, SVM, KNN y ANN (Patón 2021).

### 2.1.1 Bosque Aleatorio (RF)

El método de *RF* se basa en el algoritmo de árboles de decisión. Dicho algoritmo sirve como modelo predictivo para examinar las observaciones de un objeto en las "ramas" y deducir el valor objetivo del objeto de las "ramas". Los nodos simbolizan la segmentación de los datos. La estructura que tiene un árbol de decisión es la siguiente:

- **El nodo raíz:** El punto inicial del árbol de decisión. Este nodo representa la totalidad de los datos, la cual se parte en dos o más grupos que se puedan dividir.

- **Nodo hoja:** Estos nodos representan el resultado final.

Los sub-nodos pueden crecer del nodo raíz formando “ramas” las cuales se basan en parámetros específicos. Se pueden “cortar” las “ramas” de un árbol que se consideren redundantes para llegar a un resultado más rápido. Para elegir el mejor atributo dentro de cada nodo dos métodos son los más utilizados: La entropía y la impureza de Gini. Estos dos métodos ayudan a evaluar cada condición y lo bien que clasifican las muestras en una clase.

**Entropía**: La entropía es una cantidad que mide el nivel de desorden de los valores de la muestra. Puede tener valores entre cero y uno. La entropía aumenta si las muestras se dividen de manera equitativa en las clases, por lo que para elegir la mejor característica para dividir se elige la que tenga menor entropía. La fórmula de entropía se describe con la ecuación (1).

$$\boldsymbol{S = \sum_{c \epsilon c} p(c) log_2 p(c)} \quad \textbf{(1)}$$

Donde:

- *S* representa el conjunto de datos en el nodo del que se obtiene la entropía.
- *c* representa las clases del grupo *S*.
- *p(c)* representa la proporción de puntos de datos que pertenecen a la clase *c* con respecto al número total de puntos de datos del conjunto *S*.

**Impureza de Gini**: Este método da como resultado cual es la probabilidad de clasificar incorrectamente un dato aleatorio en el conjunto de datos si se etiquetara de acuerdo con la distribución de clases del conjunto

de datos. Es decir, si su valor es cero entonces seguro que pertenece a una clase. Se obtiene mediante la ecuación (2).

$$H(S) = -\sum_{i=1}^{k} p_i log_2(p_i) \quad \textbf{(2)}$$

Donde:

- *H* es la impureza de Gini.
- *S* representa el conjunto de datos en el nodo.
- *k* representa las clases del grupo *S*.
- *pi* es la proporción de elementos de clase *i* en el nodo.

El método de *RF* utiliza varios árboles de decisión para reducir la correlación entre las características de los datos, esta diferenciación se observa en la Figura 1. Esto se logra al elegir un número aleatorio de casos y de características.

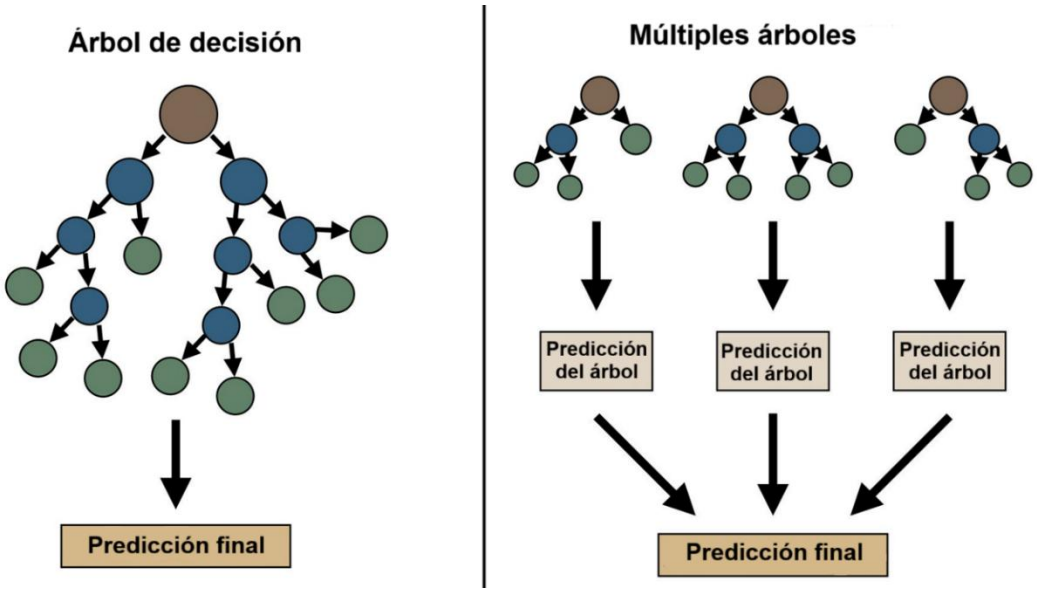



*Figura 1. Representación gráfica entre un Decision Tree y un RF*

En un inicio se toman partes iguales de los datos originales. Después se toman características de manera aleatoria para construir los árboles de decisión. Gracias a esto se logra que los distintos árboles de decisión tengan una baja correlación entre cada uno, mitigando el potencial sobre ajustamiento del modelo (Salman 2024). Los hiper parámetros que pueden ser modificados para ajustar el RF dependiendo de cada caso se pueden apreciar en la Tabla 1.

*Tabla 1. Hiper parámetros del RF*

| Hiper-parámetro | Descripción |
|---|---|
| **Número de árboles** | Número de árboles que componen el bosque. Cada árbol se entrena sobre una muestra aleatoria del conjunto de datos. Al aumentar este valor, el modelo se vuelve más robusto. |
| **Profundidad** | Profundidad máxima que puede alcanzar cada árbol. Limita cuántas divisiones pueden hacerse desde la raíz hasta las hojas. |
| **Número de características** | Número de variables que se seleccionan aleatoriamente en cada división de un nodo. |
| **Muestra por nodo** | Número mínimo de muestras requeridas en un nodo para poder dividirlo en dos ramas. Si un nodo tiene menos muestras que este valor, se convierte en hoja. |

### 2.1.2 Máquinas de Soporte Vectorial (SVM)

Las SVM se consideran como clasificadores lineales, también llamados hiperplanos. La idea del método es encontrar un hiperplano que equidista los ejemplos de cada clase, para obtener lo que se denomina margen máximo a cada lado del hiperplano. Los ejemplos de un hiperplano no óptimo y uno óptimo se aprecian en las imágenes 2 y 3.

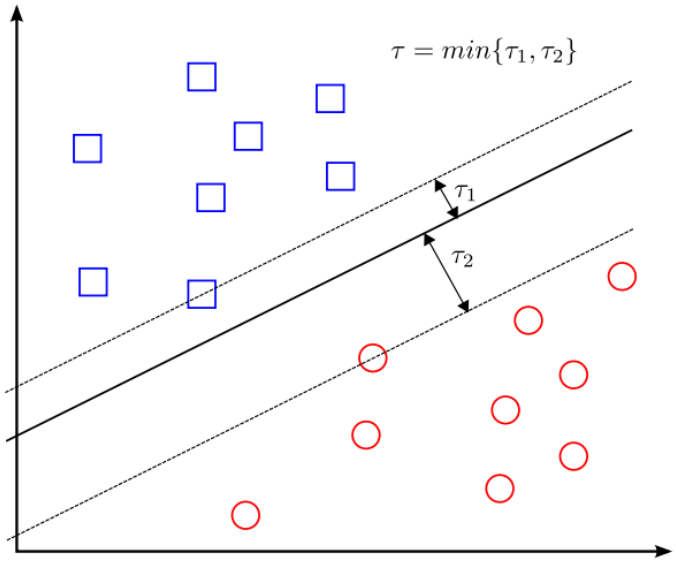



*Figura 2. Hiperplano no óptimo. Fuente: Suarez E. J. 2014*

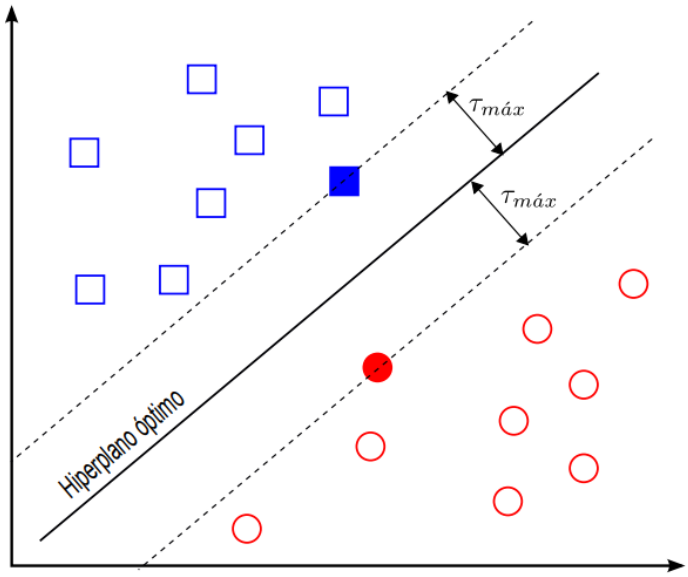



*Figura 3. Hiperplano óptimo. Fuente: Suarez E. J. 2014*

La Figura 4 muestra cómo funciona un kernel, mientras que su matemática se puede describir con la ecuación (3). Teniendo en cuenta que la distancia de $\tau_{max}$ debe cumplir con la ecuación (4) para que la distancia entre ambos grupos sea óptima.

$$w^T x + b \tag{3}$$

$$\min_{w,b} \frac{1}{2} \|w\|^2 \tag{4}$$

Donde:

- *w* es el vector de pesos definido por el modelo.
- *x* es el conjunto de características.
- *b* es el sesgo.

En el ejemplo anterior se muestra un conjunto de datos que puede ser dividido por un hiperplano lineal, pero hay casos en donde la distribución de las clases no es lineal, por lo cual no se pueden separar por una función lineal. En el caso de los datos con distribuciones no lineales se separan utilizando una función de transformación sobre los datos, con lo cual se puede aplicar el hiperplano sobre los datos. A la función de transformación que se utiliza para transformar los datos se le llama kernel (Suarez 2014). La descripción gráfica de cómo actúa un kernel se ve en la Figura 4. Las fórmulas matemáticas de los 3 tipos de kernel más usados, el lineal, el polinómico y el Gaussiano, se aprecian en las ecuaciones (5) a (7) respectivamente.

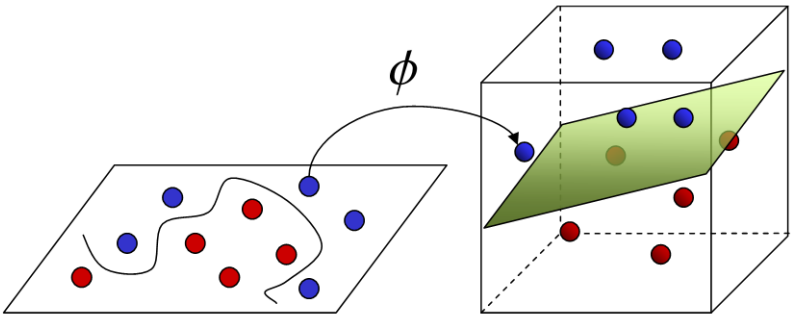



*Figura 4. Abhishek. 2024. Utilización de un kernel. Medium. (https://medium.com/@abhishekjainindore24/svm-kernels-and-its-type-dfc3d5f2dcd8)*

$$K(x_i, x_j) = x_i^T x_j \quad (5)$$

$$K(x_i, x_j) = (x_i^T x_j + c)^d \quad (6)$$

$$K(x_i, x_j) = e^{(-\frac{\|x_i - x_j\|^2}{2\sigma^2})} \quad (7)$$

Donde:

- *K* es la función del kernel.
- Donde cada $x_i/x_j$ es un valor de características
- $\sigma$ es la anchura de la campana Gaussiana.

Los hiper parámetros que pueden ser modificados para ajustar el SVM dependiendo de cada caso se pueden apreciar en la Tabla 2.

*Tabla 2. Hiper parámetros de las SVM*

| Hiper-parámetro | Descripción |
|---|---|
| **C** | Parámetro de penalización. Controla el equilibrio entre maximizar el margen y minimizar el error de clasificación. |
| **Kernel** | Define el tipo de kernel a utilizar en el caso de que los datos no se puedan separar por una función lineal. |

### 2.1.3 K Vecinos Más Cercanos (KNN)

Este algoritmo es un método que se basa en la asunción de que una instancia tiene las mismas características que los k elementos de instancias que están más cerca de esta. Este algoritmo se aprecia de manera gráfica en la Figura 5.

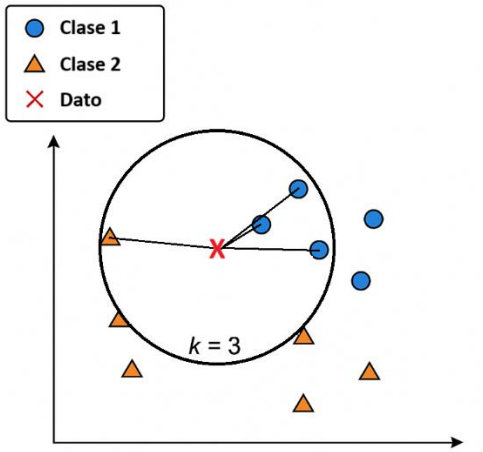



*Figura 5. Representación gráfica del algoritmo KNN. Fuente: Autoría propia.*

Para determinar cuáles son los elementos $x_i$ que están más cerca del elemento $x$ que se quiere catalogar, se mide la distancia $d(x, x_i)$ entre las $n$ características, por lo que se suele utilizar la distancia euclídea, que se muestra en la ecuación (8) (Patón 2021).

$$d(x, x_i) = \sqrt{\sum_{k=1}^{n} (x_k - x_{ik})^2} \tag{8}$$

Entre todos los algoritmos que se utilizan en el presente trabajo, el más sencillo y con menor costo computacional es el método de KNN. Incluso, su único hiper parámetro es el valor de *k*.

### 2.1.4 Redes Neuronales Artificiales – Perceptrón Multicapa (ANN – MLP)

Las ANN son modelos de ML que recrean modelos simples del cerebro humano para resolver problemas complejos. Entre los diferentes tipos de ANN, las más utilizadas por su versatilidad son las ANN perceptrón multicapa (del inglés Multi Layer Perceptron o MLP) las cuales consisten en múltiples

capas de neuronas interconectadas. Estas redes están compuestas por neuronas o perceptrones simples como se aprecia en la Figura 6.

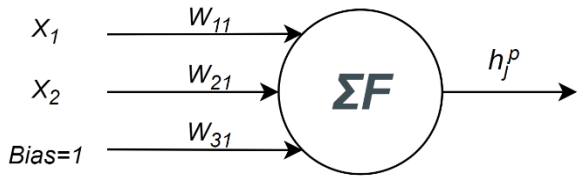



*Figura 6. Neurona o perceptrón simple. Fuente: Autoría propia.*

Donde $X_1$ y $X_2$ son valores de entrada, el $bias$ es un valor necesario para evitar que la función utilizada dentro del perceptrón se indetermine, los valores de $W_{11}$, $W_{21}$ y $W_{31}$ son los pesos por los que se multiplican las entradas para obtener el valor correcto de la salida $h_j^P$, el valor de *i* se refiere al número de neurona, *j* es el índice del número del peso de la neurona, *p* es el índice del entrenamiento y *n* es el número de neuronas de la capa anterior. Los pesos son los valores que el modelo estará cambiando constantemente para obtener el valor de salida. Dentro del perceptrón simple se suman las entradas y se multiplican por los pesos para obtener la función de excitación $S$, tal como se muestra en la ecuación (9).

$$S = \sum_{j}^{n} w_{ij} * x_i + w_{nj} * bias \tag{9}$$

Una vez obtenida la función de excitación se utiliza este resultado en la función de transferencia para obtener el resultado $h_j^P$. Las funciones que se suelen utilizar dependen de la distribución de los datos, aunque el programador puede elegir la que más le convenga según su criterio, estas funciones se observan en las ecuaciones (10) a (14).

- Lineal

$$h_j^p = S_j^p \tag{10}$$

- Escalón

$$h_j^p = \begin{cases} 0 & S_j^p < 0 \\ 1 & S_j^p \geq 0 \end{cases} \tag{11}$$

- Escalón simétrico

$$h_j^p = \begin{cases} -1 & S_j^p < 0 \\ 1 & S_j^p \geq 0 \end{cases} \tag{12}$$

- Sigmoide

$$h_j^p = \frac{1}{1 + e^{-S_j^p}} \tag{13}$$

- Tangente hiperbólica

$$h_j^p = \frac{e^{S_j^p} - e^{S_j^p}}{e^{S_j^p} + e^{S_j^p}} \tag{14}$$

Una red neuronal está compuesta por varias neuronas, donde las neuronas de la primera capa son la fila de las neuronas de entrada donde llegan los valores de entrada de la red, la última capa son la o las neuronas de salida que proporcionan él o los resultados y las capas ocultas son las capas intermedias que no están conectadas con ninguna entrada o ninguna salida (Matich 2001). En la figura de la Figura 7 se aprecia mejor la estructura de una red neuronal.

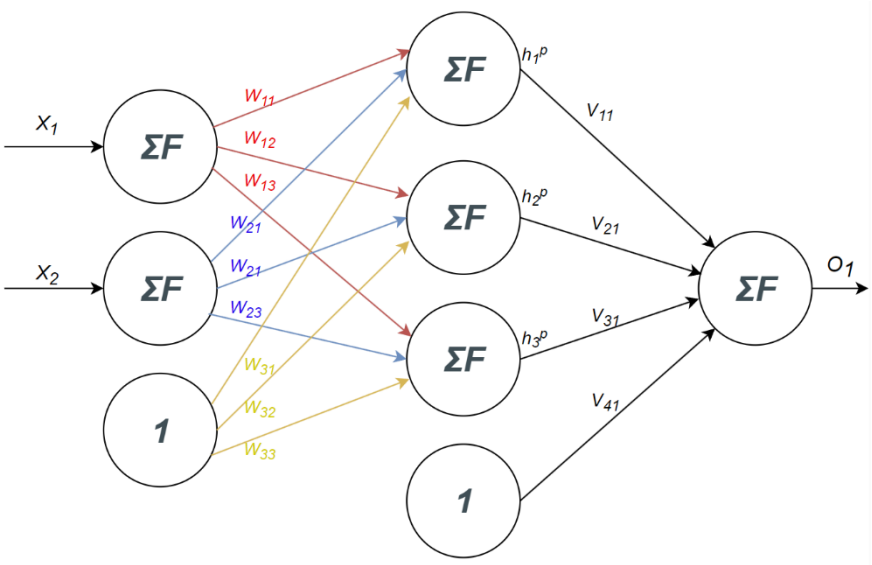



*Figura 7. Red neuronal. Fuente: Autoría propia.*

Los hiper parámetros principales que se pueden modificar en una ANN-MLP para adaptarla dependiendo de la aplicación se pueden apreciar en la Tabla 3.

*Tabla 3. Hiper parámetros de la ANN-MLP*

| Hiper-parámetro | Descripción |
|---|---|
| **Número de capas ocultas** | Indica cuántas capas intermedias hay entre la entrada y la salida. Una capa oculta puede aproximar muchas funciones, pero más capas permiten representar patrones más complejos. |
| **Número de neuronas por capa** | Determina cuántos nodos tiene cada capa oculta. |
| **Función de activación** | Introduce no linealidad en las neuronas, permitiendo modelar relaciones complejas. |
| **Tasa de aprendizaje** | Controla el tamaño de los pasos en la actualización de pesos durante el entrenamiento. |
| **Número de épocas** | Define cuántas veces se recorre el conjunto de entrenamiento completo. |

## 2.2 Métodos para visualización

Hay un problema con los datos que tienen alta dimensionalidad, como los que tienen múltiples variables, el cual es que no se pueden apreciar gráficamente. Por ello se suelen utilizar algoritmos para reducir la dimensionalidad de los datos sin perder información en el proceso y representarlos visualmente de una forma simple. Se utilizan 2 métodos en el presente trabajo para visualizar los datos, los cuales son los más comunes en este tipo de caso.

### 2.2.1 Análisis Discriminante Lineal (LDA)

Es el algoritmo más utilizado para la reducción de dimensionalidad. Su objetivo es hacer una proyección de los datos originales dentro de un nuevo subespacio de menor dimensión. Debido a que el algoritmo clasifica los datos sobre una separación, antes se maximiza dicha separación con el propósito de evitar el sobre ajuste (Serrano 2020). La descripción gráfica de cómo actúa la LDA y su comparación con la PCA se muestra en la Figura 8.

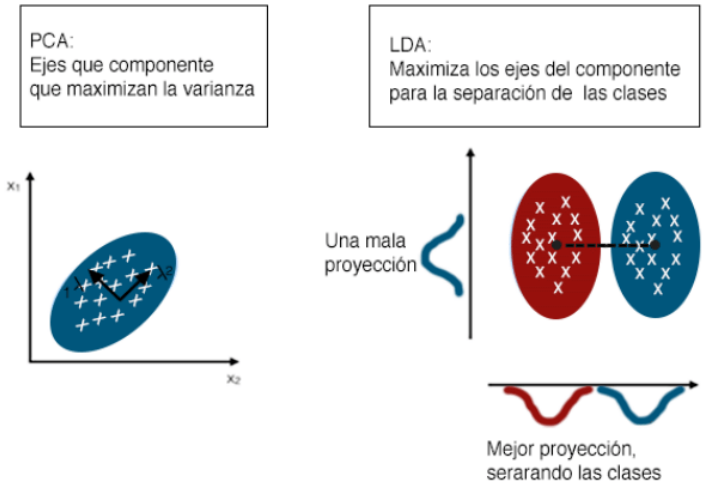



*Figura 8. Técnica de LDA presentada gráficamente. Fuente: Aguayo et al. 2018.*

La técnica de LDA busca encontrar un vector de proyección que maximice la razón entre la varianza entre clases y la varianza intraclase. Por lo tanto, es necesario obtener estos valores, para la varianza inter-clase $S_w$ se usa la ecuación (15) mientras que para la varianza entre clases $S_B$ se usa la ecuación (16). Una vez obtenidas se utilizan estos valores para calcular el vector de proyección $w$ como se aprecia en la ecuación (17), donde el valor optimo de $w$ se obtiene al cumplir con la ecuación (18).

$$S_w = \sum_{k=1}^{K} \sum_{x_i \epsilon C_k} (x_i - \mu_k)(x_i - \mu_k)^T \tag{15}$$

$$S_B = \sum_{k=1}^{K} N_k(\mu_k - \mu)(\mu_k - \mu)^T \tag{16}$$

$$J(w) = \frac{S_B w^T w}{S_w w^T w} \tag{17}$$

$$S_w^{-1} S_B w = \lambda w \tag{18}$$

Donde:

- *x* es el vector de características de entrada.
- *K* es la clase.
- *μ* es la media de los datos.
- $N_k$ es el número de muestras de la clase *k*.

### 2.2.2 Análisis de Componentes Principales (PCA)

Es un algoritmo no supervisado con diferentes aplicaciones. Al ser un algoritmo no supervisado, tiende a buscar patrones al no diferenciar entre las clases de los datos. Por ello el PCA encuentra las correlaciones entre las variables para poder separar los datos. El PCA encuentra las direcciones de máxima varianza entre datos de alta dimensionalidad y los proyecta sobre un

espacio de dimensionalidad menor, todo mientras conserva la mayor cantidad de información (Serrano 2020). Este método se visualiza en la Figura 9.

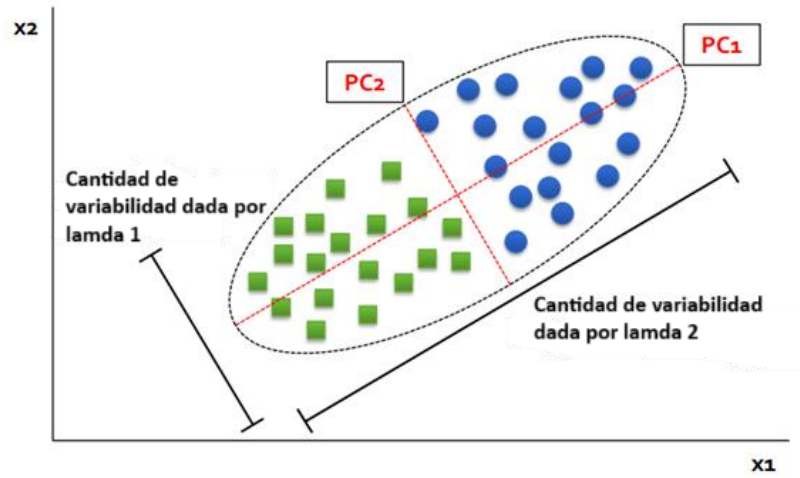



*Figura 9. Técnica de PCA presentada gráficamente. Creación propia.*

Para aplicar el método de PCA sobre la matriz de datos $X$, es necesario obtener la matriz de covarianza $S$ con la ecuación (19). Con la matriz de covarianza se pueden obtener los valores propios $\lambda_i$ y los vectores propios $v_i$ como se aprecia en la ecuación (20). Teniendo los vectores propios se eligen los primeros 3 para una representación en 3 dimensiones o los primeros 2 para una representación en 2 dimensiones, en ambos casos se aplica la ecuación (21).

$$S = \frac{1}{n-1} X^T X \tag{19}$$

$$S v_i = \lambda_i v_i \tag{20}$$

$$Z = XV \tag{21}$$

Donde:

- $Z$ es la matriz de datos ya transformados.

- $V$ son los vectores propios escogidos.

## 2.3 Análisis de varianza (ANOVA)

El método ANOVA (por sus siglas en inglés: Analysis of Variance, Análisis de Varianza) es una técnica estadística utilizada para comparar las medias de dos o más grupos y determinar si al menos una de ellas es significativamente diferente. Este método permite evaluar el efecto de uno o varios factores categóricos al comparar las medias de una variable de respuesta continua entre los distintos niveles de esos factores. Bajo la hipótesis nula, se asume que todas las medias poblacionales son iguales; en cambio, la hipótesis alternativa plantea que al menos una de ellas difiere.

Para aplicar un ANOVA, se requiere una variable dependiente continua y al menos un factor categórico con dos o más niveles. Aunque el análisis presupone que los datos provienen de distribuciones normales con varianzas similares entre los grupos, el método es relativamente robusto frente a violaciones leves de la normalidad. Sin embargo, su desempeño puede verse afectado si las distribuciones son muy asimétricas o si existe una gran diferencia en las varianzas. En tales casos, aplicar transformaciones a los datos puede ayudar a cumplir estos supuestos. La forma en que se analiza la pertenencia a estos grupos se puede apreciar en la Figura 10.

El método de ANOVA nos puede proporcionar un valor-$p$ (*p-value*) que representa la probabilidad de obtener los resultados observados (o más extremos), suponiendo que la hipótesis nula es cierta. Un valor-$p$ pequeño (usualmente < 0.05) indica que es poco probable que las diferencias en las medias se deban al azar. En ese caso, se rechaza la hipótesis nula y se

concluye que al menos un grupo tiene una media significativamente distinta (Otero 2025).

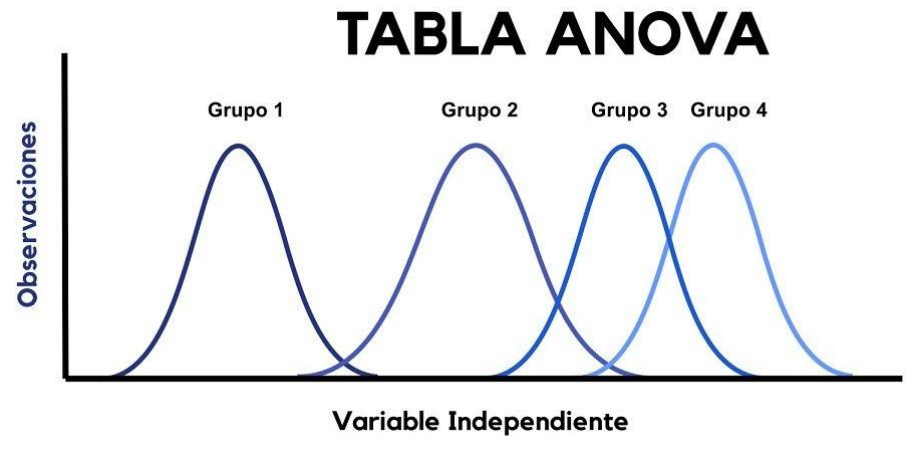



*Figura 10. 2020. Visualización de cómo trabaja ANOVA [Gráfico]. Estamatica. https://estamatica.net/tabla-anova-con-spss/*

## 2.4 Desempeño de los algoritmos

### 2.4.1 Indicadores estadísticos

Los indicadores estadísticos son valores que permiten condensar la información de un gran grupo de datos. Gracias a esto podemos conocer las tendencias, distribuciones, sesgos, modas, dispersiones, entre otras, de un grupo sin necesidad de conocer los valores de cada uno de los individuos de dicho grupo. Los indicadores estadísticos han demostrado ser eficientes a la hora de estimar modelos o propiedades de una señal o un conjunto de datos. Una de sus características es que implican una baja carga computacional lo que permite que estos datos se puedan obtener rápidamente. Además, se han utilizado estos indicadores para entrenar métodos de ML en diversos campos, desde el campo de la economía, para la optimización de portafolios de inversión (Mazo 2025), hasta el campo de la ingeniería, para la detección de fallas en motores (García 2021). Debido a la naturaleza de los datos, hay indicadores estadísticos que no se pueden aplicar o que proporcionan

información repetida, o redundante, que podría reducir sus propiedades discriminatorias o evitar encontrar perfiles o patrones discriminantes. Por ello solo se utilizan los indicadores estadísticos que se aprecian en la Tabla 4, la cual contiene las ecuaciones (22) a (27) que muestran cómo se obtiene cada indicador.

*Tabla 4. Fórmulas de indicadores estadísticos*

| Indicador estadístico | Descripción | Ecuación | |
|---|---|---|---|
| **Media ($\bar{x}$)** | Valor promedio de los datos. | $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ | (22) |
| **Varianza ($\sigma^2$)** | Promedio de los valores cuadrados de la desviación de esa variable. | $\sigma^2 = \frac{1}{N} * \sum_{i=1}^{N} (x_i + \bar{x})^2$ | (23) |
| **Desviación estándar ($\sigma$)** | Dispersión de los datos en base su media. | $\sigma = \sqrt{\frac{1}{N} * \sum_{i=1}^{N} (x_i + \bar{x})^2}$ | (24) |
| **Valor máximo** | Valor máximo del conjunto de datos. | $x_{max} = max\{x_1, x_2, x_3, \dots x_n\}$ | (25) |
| **Skewness ($S_k$)** | Medida de la asimetría de la probabilidad de distribución de los datos. | $Sk = \frac{1}{N} * \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{\sigma}\right)^3$ | (26) |
| **Kurtosis ($k_u$)** | Indicador del grado de nitidez de la distribución de probabilidad. | $Ku = \frac{1}{N} * \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{\sigma}\right)^4$ | (27) |

Donde $N$ es el total de datos de la muestra, $x_i$ es el i-esimo dato de la muestra, y $\bar{x}$ es la media de los datos.

### 2.4.2 Métricas de desempeño

Es importante tener métricas de rendimiento para discriminar diferentes algoritmos de ML. Dichas métricas de desempeño se consiguen aplicando una matriz de confusión para variables dicotómicas. Un ejemplo de una matriz de confusión como la que se utiliza en este trabajo es mostrado en la Tabla 5.

*Tabla 5. Matriz de confusión*

| | | **Datos referenciados** | | |
|---|---|---|---|---|
| | | Grupo 1 = Positivo | Grupo 2 = Negativo | Total |
| **Resultados de la clasificación** | Grupo 1 = Positivo | F11 = verdadero positivo | F10 = falso negativo | F1T |
| | Grupo 2 = Negativo | F01 = falso positivo | F00 = verdadero negativo | F0T |
| | | F1T | F0T | N |

En donde F1T y F0T son la cantidad total de instancias clasificadas como positivo o negativo por el algoritmo respecto a ese grupo y N es el número total de muestras. Las métricas que se obtienen con estos datos son:

- "***Accuracy***" o exactitud (en inglés) nos permite saber que tan efectivo es nuestro algoritmo en acercarse al valor real de la clasificación.
- "***Recall***" o sensibilidad (en inglés) que nos da a conocer la proporción de casos positivos que fueron correctamente clasificados.
- "***Precision***" o precisión (en inglés) nos permite saber que tan cerca están los valores conseguidos entre sí, o también la repetitividad que tiene nuestro algoritmo tiene para dar de nuevo el mismo resultado con el mismo dato.

- “**F1**” es una medida que combina “*accuracy*” y “*recall*” por lo que nos dice que tanto se acerca nuestro algoritmo a los casos reales y positivos.

En general los resultados de exactitud, sensibilidad, precisión y F1 se consideran aceptables cuando están en un rango de 0.6-0.75, buenos cuando están en un rango de 0.75-0.9 y excelentes cuando están en un rango de 0.9-1.0.

La representación gráfica de las métricas de desempeño se ilustra en la Figura 11, mientras que las ecuaciones (28) a (31) se utilizan para obtener estas métricas (Borja 2020).

$$Accuracy = \frac{f_{11} + f_{00}}{N} \tag{28}$$

$$Recall = \frac{f_{11}}{f_{11} + f_{01}} \tag{29}$$

$$Precision = \frac{f_{11}}{f_{11} + f_{10}} \tag{30}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{31}$$

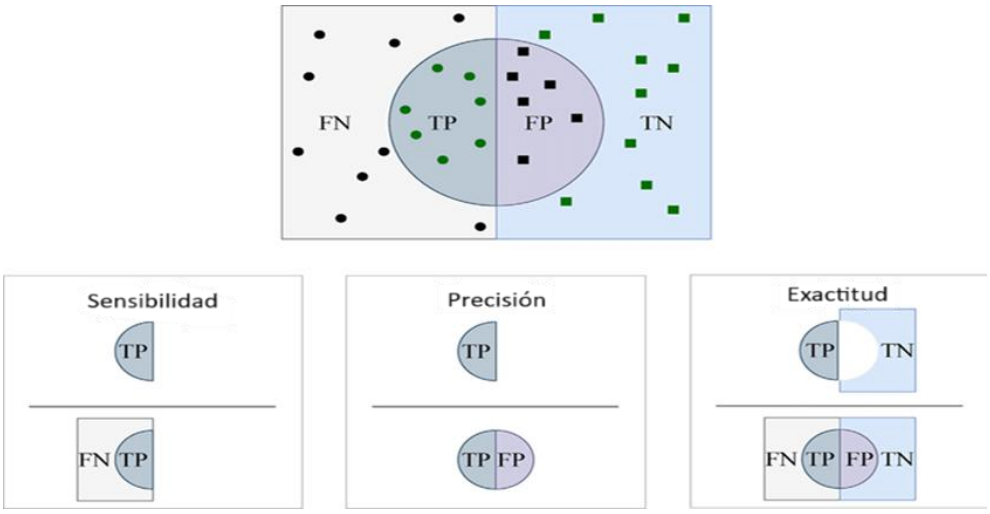



*Figura 11. Representación de exactitud, sensibilidad y precisión. Fuente: Propia*

## 2.5 Bases biológicas

A continuación, se presenta la fundamentación teórica referente a los temas de base biológica necesarios para el desarrollo de la propuesta metodológica del presente trabajo.

### 2.5.1 Microbiota

Los microorganismos son seres invisibles para el ojo humano y habitan en todos los rincones del mundo y, en todos los ecosistemas, tienen una relación muy cercana con otros seres vivos. Al grupo de microorganismos que habitan en una zona determinada se le llama microbiota y está conformada por una gran cantidad de distintas especies de hongos, bacterias y protozoarios. Muchos autores también llaman microbiota al conjunto de millones de microorganismos que habitan en una relación simbiótica con el ser humano, es decir, ambos organismos obtienen un beneficio mutuo. Estos microorganismos habitan en nuestra piel, boca, genitales y sistema gastrointestinal (Moreno 2022).

### 2.5.2 Categoría taxonómica

Por su naturaleza, el hombre tiende a clasificar los objetos, conceptos e incluso seres de su entorno, de esta manera puede agruparlos por sus características en común y diferenciarlos de otros por sus diferencias. La taxonomía se basa en los principios de la clasificación y se basa en tres procesos. El primero es la clasificación en sí, donde se busca agrupar los objetos o seres que poseen semejanzas entre sí para diferenciarlos de los que no comparten dichas semejanzas. El segundo es la nomenclatura, por el que se nombran cada uno de los grupos creados. El tercero es la identificación, por el que se clasifica un ser u objeto dentro de las categorías o grupos creados o, por el contrario, se cataloga como inclasificable dentro de los grupos preexistentes.

Los grupos creados dentro del contexto biológico se suelen llamar taxones. Estos taxones tienen los siguientes nombres, de rango superior a rango inferior: dominio, reino, filo, clase, orden, familia, género y especie. En ciertas ocasiones hay un taxon inferior denominado subespecie. Todo taxon superior está conformado por varios taxones semejantes del siguiente nivel inferior (Prats et al. 2002). Los niveles taxonómicos por nivel se pueden apreciar gráficamente en la Figura 12.

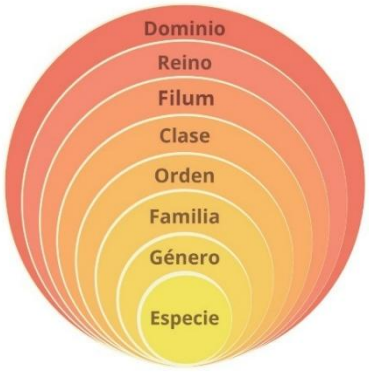



*Figura 12. Calcáneo, M. G. I. y de la Cueva, B. L. (2021). Categorías taxonómicas. En Características generales de los dominios y los reinos. Portal Académico del CCH, UNAM.*

*https://portalacademico.cch.unam.mx/biologia2/caracteristicas-generales-dominios-y-reinos/categorias-taxonomicas*

### 2.5.3 Microbiota intestinal

La mayor parte de la microbiota humana habita en el intestino grueso, donde el 90% de las especies pertenecen a los filos *Bacteroidetes* y *Firmicutes*, mientras que el 10% restante pertenece principalmente a los filos *Proteobacterias*, *Actinobacterias*, *Fusobacteria* y *Verrucomicrobia*.

La microbiota en el intestino cumple con funciones necesarias como la fermentación de carbohidratos complejos y la producción de vitaminas del grupo B y la vitamina K e incluso regula el almacenamiento de lípidos. La microbiota también tiene un papel importante en el sistema inmune al ayudar en la creación de la barrera física que protege el intestino y en la producción de estímulos para la regulación de células del sistema inmune.

La microbiota puede sufrir cambios sustanciales debido a los cambios de alimentación, las enfermedades, el uso de medicamentos, etc., pero esta composición tiende a volver a su estado inicial después de dichas fluctuaciones. La composición de la microbiota en un estado normal es distinta para cada individuo por múltiples factores, ya sea por el tipo de parto en el que el individuo fue concebido o la exposición a distintas bacterias a lo largo de la vida, pero los hábitos de vida de la persona como la dieta, el ejercicio y la exposición al estrés, etc., también tienen un gran peso en el balance entre las bacterias benéficas y las bacterias patógenas que existen en el intestino (Moreno 2019).

### 2.5.4 Disbiosis

La disbiosis se puede describir como una alteración compositiva y funcional de la microbiota intestinal causada por un factor externo o del individuo. Aunque la microbiota tiende a adaptarse a los cambios que se presentan, como pueden ser los picos de estrés, la mala alimentación, el uso de medicamentos o el sufrir de alguna enfermedad, cuando la afectación ha desaparecido y la composición microbiana no ha regresado a su estado normal y persiste crónicamente tiende a tener consecuencias perjudiciales para el huésped. La disbiosis puede tener 3 variantes generales que se diferencian por su característica principal, aunque no son excluyentes entre sí:

- **Incremento de patobiontes**. Los patobiontes son bacterias que tienden a causar patologías, estos patobiontes pueden existir en el intestino del individuo en bajas cantidades, pero una anomalía puede ocasionar que se multipliquen hasta puntos donde causan afectaciones al cuerpo.
- **Disminución o pérdida de comensales**. Se les suele llamar comensales a las bacterias “benéficas” o que proporcionan una función útil para el cuerpo humano. Las bacterias patógenas suelen causar una pérdida de comensales por la reducción de la proliferación, la competencia por recursos o la muerte microbiana. La restauración de las bacterias abolidas o de sus metabolitos tiene potencial de revertir los fenotipos causados por la disbiosis.
- **Pérdida de la diversidad microbiana**. La disbiosis relacionada con una enfermedad tiende a reducir la diversidad microbiana inicial.

En muchos casos una disbiosis se presenta a la par de una enfermedad o de afectaciones negativas sobre el individuo, pero no siempre existe una relación causalidad directa entre la disbiosis y la enfermedad o viceversa. Pero estudios recientes demuestran que hay un perfil de la

disbiosis asociado con las enfermedades gastrointestinales. Las patologías intestinales pueden caracterizarse por un descenso de la diversidad microbiana, un incremento de bacterias patobiontes y/o una disminución de las bacterias comensales (Moreno 2022).

Normalmente para el análisis de disbiosis se suele prestar atención a los filos de bacterias ya que permiten ver afectaciones a gran escala sin tener que revisar el siguiente taxon inferior, lo cual requiere de un análisis más profundo. Debido a que la base de datos no es tan robusta como para mostrar todas las clases de bacterias, en el trabajo presente se opta por analizar los filos de bacterias.

## 2.6 Índices de diversidad

Se les conoce como índices de diversidad a la representación matemática de la diversidad biológica de una zona determinada, basándose en la cantidad de especies y el número de individuos de cada especie. Cada índice tiene una fórmula distinta debido a que cada autor basó su fórmula en distintos criterios, por lo cual se suelen utilizar varias fórmulas para que se puedan contrastar resultados con otros estudios.

Los índices de diversidad más utilizados se observan en las ecuaciones (32) a (36). De los cuales las cinco se pueden utilizar para el mismo cometido, y deben de ser utilizadas la mayoría para evitar el que no se puedan contrastar resultados, pero, por comodidad, las ecuaciones (32), (35) y (36) se utilizan para entornos con muchas especies y las ecuaciones (33) a (34) cuando hay una gran cantidad de especies.

- Riqueza de especies:

$$D_{rich} = S = \sum p_i^0 \quad (32)$$

- Índice de Shannon:

$$H_{Shannon} = -\sum p_i * log_b(p_i) \quad (33)$$

- Exponencial del índice de Shannon:

$$D_{\exp Shannon} = b^{H_{Shannon}} \quad (34)$$

- Índice de Gini-Simpson:

$$H_{Gini-Simpson} = 1 - \sum p_i^2 \quad (35)$$

- Inverso del índice de Gini-Simpson:

$$D_{inv\ Gini-Simpson} = \frac{1}{1 - H_{Gini-Simpson}} = \frac{1}{\sum p_i^2} \quad (36)$$

Donde $pi$ es la abundancia relativa de la especie $i$, es decir la abundancia de la especie $i$ dividida entre la suma de las abundancias de las $S$ especies de la comunidad analizada, y $b$ es la base del logaritmo con el que se está trabajando (González 2012).

# CAPITULO 3: METODOLOGÍA

A continuación, se presenta la metodología propuesta basada en la gestión de datos mediante técnicas de IA en una estructura de ML para el diagnóstico del SII para pacientes usando bases de datos en línea, explicando en detalle los pasos necesarios para su aplicación. En la Figura 13 se aprecia el diagrama de bloques general del trabajo propuesto, en el cual se aprecian tres etapas generales: 1) selección de las bases de datos de bacterias, 2) obtención de los dos archivos principales con información de las abundancias y filos de bacterias, y 3) desarrollo de la herramienta de diagnóstico para el SII.

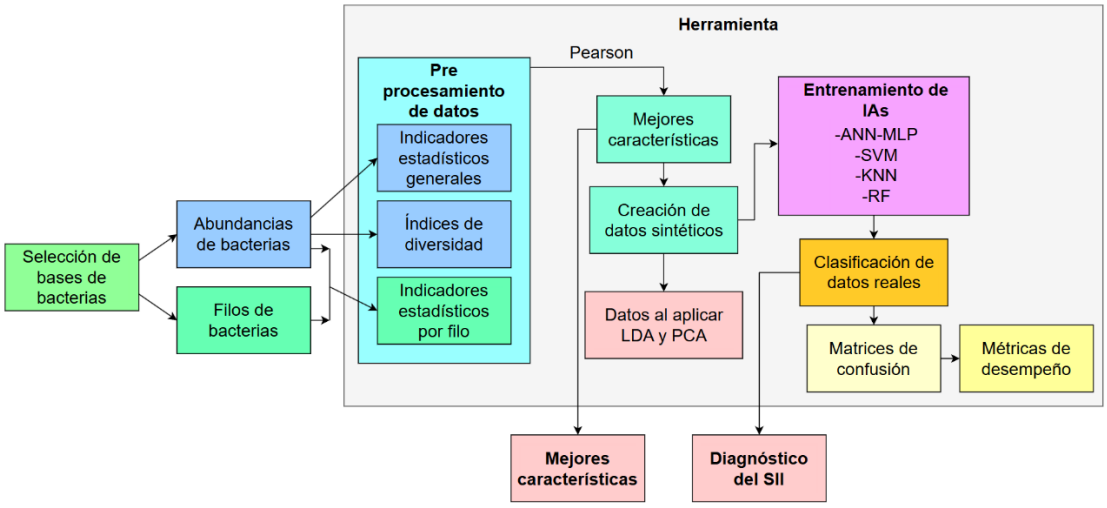



*Figura 13. Diagrama de bloques de la metodología basada en la gestión de datos mediante técnicas de IA en una estructura de ML para el diagnóstico del SII (Autoría propia)*

Dentro de la tercera etapa que consiste en el desarrollo de la herramienta para el diagnóstico del SII se tienen varias subetapas: i) realizar el preprocesamiento de los datos donde se obtienen los indicadores estadísticos generales, los índices de diversidad y los indicadores estadísticos de los filos de bacterias de la base de datos; ii) se obtienen las mejores características mediante el coeficiente de Pearson para la creación

de datos sintéticos ( esto con el fin de visualizar los datos sintéticos y los datos reales mediante LDA y PCA); iii) con los datos sintéticos se entrenan los métodos y se prueban en los datos reales para obtener los resultados y las métricas de cada uno. Los resultados visibles para el usuario son las mejores características y el diagnóstico del o de los pacientes.

## 3.1 Selección de la base de datos

Para desarrollar y entrenar las cuatro técnicas de IA que se planean utilizar es necesaria una base de datos. Por esto se elige una base de datos de microbiota disponibles en línea, la cual ha sido creada por particulares y/o escuelas que han obtenido estos datos de manera controlada y segura. El uso de estas bases de datos es libre y sus autores buscan que esta información sea utilizada por terceros mientras se otorgue el crédito a las debidas instituciones. Ninguna información personal, de identificación o sensible se encuentra en estas bases de datos, simplemente se incorporan los datos de microbiota, información de los experimentos y meta data relacionada a los hábitos de salud de la gente. La página web de donde se obtuvo la base de datos es “https://www.ebi.ac.uk/”, mostrada en la Figura 14.

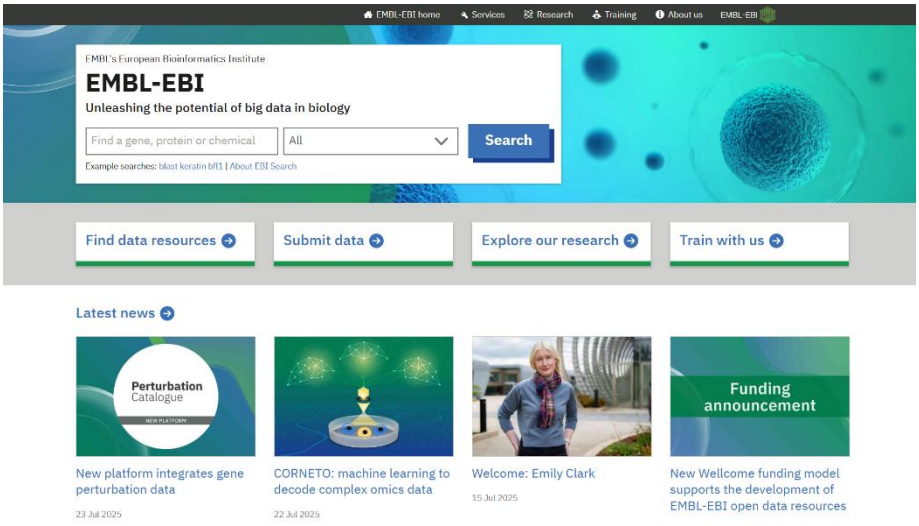



*Figura 14. Página fuente de los datos utilizados*

En esta se pueden encontrar diversos estudios, artículos y trabajos junto con los datos que utilizaron. Las bases de datos están disponibles en formatos ".sra" y ".lite.1", los cuales se pueden convertir a formato ".csv". Se optó por usar esta página debido a que proporciona todos los datos necesarios y sin procesar.

La base de datos que se decidió utilizar proviene del estudio titulado “Follow-up of faecal microbiota in IBS patients”, el cual contiene 39 estudios de microbiota intestinal de 39 personas distintas, de estos casos 30 son de personas con SII y 9 son de control. Se optó por usar esta base de datos porque cuenta con el mayor número casos que contiene un grupo de control de personas sanas y con SII.

Al transformar los archivos de la base de datos de www.ebi.ac.uk/ de su formato “.sra” y “.lite” a formato “.csv” se obtienen 3 archivos: El primero tiene los datos del proyecto como el número de serie del proyecto, el tipo de análisis del experimento, número de serie del experimento, número de identificación del paciente (que esta codificada y es usada solo para diferenciar a cada uno de los pacientes y relacionarlos con sus datos), fecha de obtención, entre otros datos que sirven para el manejo y control interno de la herramienta que contiene la base de datos pero no se utilizan en este trabajo. Además de datos no personales del paciente como sexo, dieta, antecedentes de tabaquismo, etc. Un ejemplo de los datos que se pueden encontrar en este tipo de archivos se ve a continuación, en las Tablas 6 y 7.

*Tabla 6. Datos de los experimentos y de los pacientes - a*

| Acceso_análisis | Tipo_análisis_experimento | Momento_análisis | Modelo_instrumento_análisis | Fecha_recolección_muestra | Característica_medio_muestra |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| **MGYA00000383** | Amplicon | 2012-08-29T00:00:00 | 454 GS FLX Titanium | 06/07/2010 | human-associated habitat |
| **MGYA00000399** | Metatranscriptomic | 2012-08-29T00:00:00 | 454 GS FLX Titanium | 03/08/2010 | human-associated habitat |
| **MGYA00000403** | Metatranscriptomic | 2012-08-29T00:00:00 | 454 GS FLX Titanium | 13/07/2010 | human-associated habitat |
| **MGYA00000404** | metagenomic | 2012-08-29T00:00:00 | 454 GS FLX Titanium | 30/06/2010 | human-associated habitat |

*Tabla 7. Datos de los experimentos y de los pacientes - b*

| Tipo_material_muestra | Sexo_paciente | Edad_paciente | Indice_masa_corporal_paciente | Dieta_paciente | Patología_paciente |
|---|---|---|---|---|---|
| **Feces** | male | 66 | 31.7 | mediterranean diet | None |
| **Feces** | female | 62 | 27.5 | mediterranean diet | irritable bowel syndrome, diarrhoea subtype |
| **Feces** | male | 22 | 30.4 | mediterranean diet | irritable bowel syndrome, diarrhoea subtype |
| **Feces** | female | 21 | 30.4 | mediterranean diet | irritable bowel syndrome, diarrhoea subtype |

Es necesario tener los datos de entrenamiento ordenados y homogeneizados para facilitar su uso, por ello, de la base de datos elegida se toman los datos necesarios correspondientes a las personas que sufren SII y a un grupo de gente saludable como grupo de control. De dichos datos se dejarán las abundancias de bacterias y de los nombres de las especies de bacterias.

En el segundo archivo se encuentran las abundancias de cada microorganismo de todos los pacientes de la base de datos, diferenciando a cada paciente por su código de identificación y diferenciando a cada microorganismo por un código numérico, mientras que la cantidad de abundancias de bacterias se representan con números enteros, siendo la ausencia de una especie marcada con un 0. Un extracto del archivo con algunos de los datos que se pueden encontrar se aprecia en la Tabla 8, donde se muestra la información de solamente 6 pacientes. Entonces, se observa que las columnas de la tabla presentan los códigos de identificación del paciente, mientras que los renglones de la tabla indican los códigos numéricos del microorganismo. Por ejemplo, un paciente tiene el código de identificación “MGYA00637743” y el código numérico del microorganismo es “104718” con una abundancia de bacterias de “0”, es decir ausencia de bacterias.

*Tabla 8. Abundancias de microorganismos por especie en código*

| | MGYA00637743 | MGYA00637746 | MGYA00637749 | MGYA00637752 | MGYA00637755 | MGYA00637758 |
|---|---|---|---|---|---|---|
| **104718** | 0 | 1 | 0 | 0 | 0 | 0 |
| **25772** | 0 | 0 | 0 | 0 | 1 | 0 |
| **212535** | 0 | 0 | 0 | 0 | 2 | 0 |
| **11297** | 0 | 0 | 0 | 0 | 0 | 0 |
| **48212** | 0 | 0 | 0 | 0 | 0 | 0 |
| **139706** | 0 | 0 | 0 | 0 | 0 | 0 |

El tercer archivo relaciona los códigos numéricos de cada microorganismo con el reino, filo, clase, orden, familia, género y especie al que pertenecen. Adicionalmente, la leyenda “NA” se coloca cuando no se ha identificado el reino, filo, clase, orden, familia, género o especie de un microorganismo debido a que no se ha catalogado anteriormente dentro de

la base de datos del laboratorio que realizó el experimento. Nuevamente, un extracto de los datos que se pueden encontrar en este archivo se aprecia en la Tabla 9, donde los renglones indican el código numérico de los microorganismos, mientras que las columnas indican reino, filo, clase, orden, familia, género o especie, correspondiente al código numérico.

*Tabla 9. Desglose del código de especies de microorganismos*

| | Reino | Filo | Clase | Orden | Familia | Genero | Especie |
|---|---|---|---|---|---|---|---|
| **104718** | Fungi | Ascomycota | Saccharomycetes | Saccharomycetales | Dipodascaceae | Yarrowia | NA |
| **25772** | Fungi | Ascomycota | Saccharomycetes | Saccharomycetales | Saccharomycetaceae | Saccharomyces | NA |
| **212535** | Fungi | Ascomycota | Saccharomycetes | Saccharomycetales | Saccharomycetaceae | NA | NA |
| **11297** | Fungi | Ascomycota | Saccharomycetes | Saccharomycetales | NA | NA | NA |
| **48212** | Fungi | Mucoromycota | Mucoromycetes | Mucorales | NA | NA | NA |
| **139706** | Metazoa | Chordata | Mammalia | NA | NA | NA | NA |
| **165034** | Metazoa | NA | NA | NA | NA | NA | NA |
| **73657** | Viridiplantae | Streptophyta | Magnoliopsida | Fabales | Fabaceae | Ammopiptanthus | Ammopiptanthus_mongolicus |
| **94865** | Viridiplantae | Streptophyta | Magnoliopsida | Fabales | Fabaceae | Medicago | Medicago_sativa |

## 3.2 Pre-procesamiento de datos

Para facilitar el análisis de los datos se obtendrán los valores de los 5 índices de diversidad más comunes, utilizando las ecuaciones (32) a (36), y se obtendrán indicadores estadísticos de media, mediana y varianza de todas las abundancias de bacterias usando las ecuaciones (22) a (27). Para tener un análisis más a fondo se obtienen los indicadores estadísticos de media,

varianza, desviación estándar, valor máximo, skewness y kurtosis de cada uno de los filos de bacterias de la base de datos. Un extracto ejemplo del listado de las características obtenidas se aprecia en la Tabla 10.

*Tabla 10. Datos procesados y listos para usar en el ML*

| Diversidad de especies | Índice de Shannon | Exponencial Índice de Shannon | Índice de Gini-Simpson | Inverso del Índice de Gini-Simpson | Media | Varianza | Desviación estándar | Media Actinobacteria |
|---|---|---|---|---|---|---|---|---|
| 18 | 7.454 | 1728 | -34 | 0.028571 | 0.319 | 0.389 | 0.624 | 0 |
| 21 | 16.819 | 20155392 | -59 | 0.016666 | 0.444 | 0.644 | 0.802 | 0 |
| 8 | 1.386 | 4 | -10 | 0.090909 | 0.125 | 0.139 | 0.372 | 0.5 |
| 8 | 29.986 | 1.0541E+13 | -100 | 0.009900 | 0.319 | 1.319 | 1.148 | 0.5 |

## 3.3 Creación de datos sintéticos

En muchos casos las bases de datos con las que se cuenta son pequeñas y están desbalanceadas, lo que vuelve difícil entrenar modelos de ML. Por ello se puede utilizar la base de datos original para crear datos sintéticos y aumentar la base de datos de entrenamiento. Aunque un método de ML entrenado con una base de datos real tiene mejores métricas de desempeño que uno entrenado con una base con datos sintéticos, la diferencia no tiende a ser significativa (Rajotte et al. 2022).

Debido a que la base de datos utilizada no es suficientemente grande se tomó la decisión de entrenar los métodos con datos sintéticos. Para crear los datos sintéticos se obtienen las mejores características utilizando el coeficiente de Pearson. Utilizando como criterio el valor de $p < 0.05$ lo cual indica que la característica se puede considerar como estadísticamente significativa. Teniendo las mejores características se toman los rangos de los casos de personas con SII y de las personas sanas. Después, se crean 200 datos utilizando los rangos de las mejores características de las personas

con SII y 200 datos utilizando los rangos de las mejores características de las personas sanas.

## 3.4 Visualización de los datos

Para visualizar los datos y como se diferencian el grupo de personas con SII y el grupo de personas sanas es necesario reducir el número de dimensiones de los datos sin perder información. Por lo tanto, se utilizaron las ecuaciones (15) a (18) para aplicar el método LDA y representar los datos en una dimensión, mientras que las ecuaciones (19) a (21) se emplearon para aplicar el método PCA y representar los datos en tres dimensiones.

## 3.5 Hiper-parámetros de los métodos de ML

Para elegir las características de los métodos de ML se realizaron múltiples pruebas cambiando poco a poco sus métricas para hacerlo más robusto y complejo y comparando sus resultados, optando por dejar las métricas si no había un cambio significativo respecto a las siguientes mejoras. Para el método de ANN-MLP se empezaron las pruebas con 2 capas de neuronas internas, ya que con más capas podría causar sobreajuste y con una sola el método no sería suficientemente robusto, de 5 neuronas cada una y se aumentaron el número de neuronas de ambas capas de 5 en 5. Se tomaron como características finales con las cuales se dejó de observar un cambio significativo en los resultados. La función de activación es la función “ReLU” la cual se utiliza por defecto en Python, la cual resuelve los problemas de las funciones normales como la sigmoide. La tasa de aprendizaje que utiliza el programa de Python es adaptativa por defecto ya que evita la necesidad de ir cambiando la tasa.

Para el método de SVM se compararon los resultados con 3 tipos de kernel: el lineal, el polinómico y el radial. Debido a que el kernel es la

característica principal del método de SVM y la única que se puede modificar, el mejor kernel se toma como característica final. Para la penalización *C* se optó por un valor de 10, suficientemente alto para evitar errores, pero no demasiado para causar sobreajuste.

Para el método de KNN se realizaron pruebas para encontrar el valor de *k*. Se inició con un valor de *k=2* y se aumentó de 1 en 1 hasta llegar a un valor de *k=10*. Se toma como característica final el valor de *k* con mejores resultados.

Para los parámetros de RF se cambiaron los parámetros de número de árboles con valores de 100, 300 y 500, la profundidad de los árboles con valores de 10, 20 y 30, las muestras para división de nodos con valores de 2, 5 y 10. El número de características se fijó como “sqrt”, el tipo más utilizado, es decir que la cantidad de variables a considerar en la división de cada nodo será igual al cuadrado de la cantidad de características dentro de la base de datos. Esto permite reducir la correlación entre árboles permitiendo la generalización. En cuanto los resultados dejaron de tener cambios significativos se tomaron las características como las finales. Las características finales de cada método se enlistan en la Tabla 11.

*Tabla 11. Hiper-parámetros seleccionados para cata algoritmo de ML utilizado*

| Método | Características |
|---|---|
| ANN-MLP | Número de capas internas: 2<br>Neuronas de primera capa intermedia: 15<br>Neuronas de segunda capa intermedia: 15<br>Número de iteraciones: 1000<br>Tasa de aprendizaje: Adaptable |
| SVM | Kernel: Gausiano<br>C: 10 |

| kNN | k: 5 |
|---|---|
| RF | Número de árboles: 300<br>Profundidad: 20<br>Número de características: “sqrt” = 6<br>Muestras para división de nodos: 5 |

## 3.6 Entrenamiento de los métodos de ML

Con los 400 datos sintéticos se entrenan los métodos de ANN-MLP, SVM, kNN y RF. Con los métodos entrenados se clasifican los datos reales y se obtienen las métricas de desempeño de cada uno. Los resultados que se obtienen de los 4 métodos se dividen en verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. Un ejemplo se observa en la Tabla 12.

*Tabla 12. Ejemplo de los resultados de clasificación de las IA*

| | Predicción | | |
|---|---|---|---|
| Verdadero | | Positivo | Negativo |
| | Positivo | 7 | 0 |
| | Negativo | 2 | 30 |

Se compararán los resultados de los algoritmos utilizados mediante matrices de confusión utilizando las ecuaciones (28) a (31). Esto para encontrar el método más eficiente a la hora de diagnosticar el SII y basar el software final en ese método. Un ejemplo del desempeño de las técnicas de ML que se obtiene de las matrices de confusión se aprecia en la Tabla 13.

*Tabla 13. Ejemplo de métricas de desempeño de las IA*

| | |
|---|---|
| Exactitud | 95% |
| Precisión | 100% |

| | |
|---|---|
| Sensibilidad | 78% |
| F1 | 88% |

## 3.7 Herramienta de software de diagnóstico

Se realizó una herramienta de diagnóstico del SII utilizando el lenguaje de programación de Python. Dicha herramienta contiene 3 pestañas. El usuario tiene 2 opciones: Marcar la opción de default para utilizar la base de datos que se utilizó en el presente trabajo, con el cual solo es necesario subir el archivo donde se encuentran los datos del paciente, o los pacientes, que se quieren diagnosticar o, por otra parte, utilizar una base de datos propia para el entrenamiento de los métodos de ML y con estos métodos entrenados diagnosticar el o los pacientes que se requieran. Dentro de la interfaz, el usuario puede elegir los métodos a utilizar en el análisis: ANN-MLP, SVM, kNN o RF, para hacer una comparación. En el diagrama de flujo de la Figura 15 se puede apreciar cómo está estructurada e implementada la herramienta.

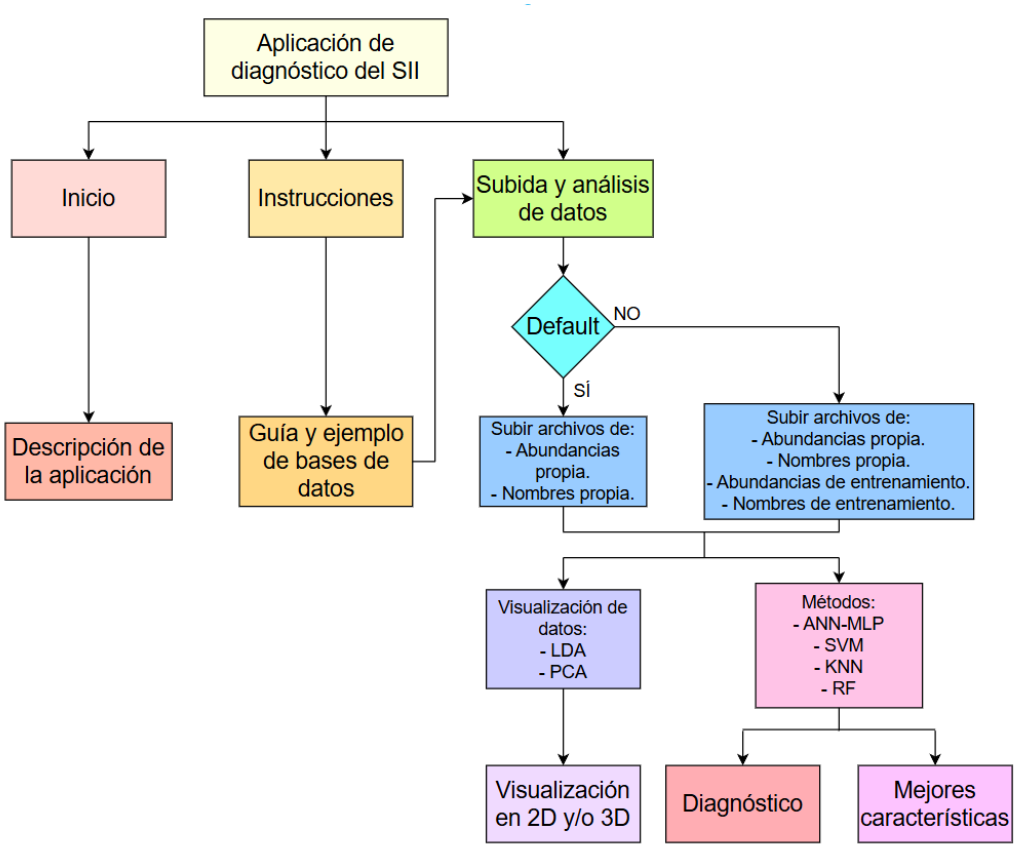



*Figura 15. Diagrama de la Herramienta de software (Autoría propia)*

# CAPITULO 4: RESULTADOS

## 4.1 Mejores características

De las 66 características que se tenían inicialmente, quedaron las características que cumplieron con el valor de $p < 0.05$ al utilizar el coeficiente de Pearson $r$, y, por tanto, se quedan las mejores 19 características que se observan en la Tabla 14.

*Tabla 14. Mejores características obtenidos con Pearson*

| Característica | *r* | *P* |
|---|---|---|
| Media Lentisphaerae | -0.7566 | 0 |
| Max Lentisphaerae | -0.7549 | 0 |
| Desviación Lentisphaerae | -0.7537 | 0 |
| Varianza Lentisphaerae | -0.6967 | 0 |
| Desviación Firmicutes | -0.5746 | 0.0002 |
| Media Firmicutes | -0.5453 | 0.0004 |
| Max Firmicutes | -0.5196 | 0.0008 |
| Varianza Firmicutes | -0.4908 | 0.0018 |
| Skewness Lentisphaerae | -0.4400 | 0.0057 |
| Media Verrucomicrobia | -0.4309 | 0.0069 |
| Kurtosis Lentisphaerae | -0.4085 | 0.0109 |
| Desviación general | -0.4037 | 0.0120 |
| Desviación Verrucomicrobia | -0.3897 | 0.0156 |

| | | |
|---|---|---|
| Max Verrucomicrobia | -0.3888 | 0.0158 |
| Kurtosis Proteobacteria | -0.3476 | 0.0325 |
| Media general | -0.3442 | 0.0343 |
| Skewness Proteobacteria | -0.3418 | 0.0357 |
| Índice de Shannon | 0.3385 | 0.0377 |
| Riqueza de especies | -0.3219 | 0.0488 |

El valor de $p < 0.05$ indica que las 19 características presentadas son estadísticamente significativas, mientras que los valores de $r$ reflejan la fuerza de la correlación entre dichas características y el diagnóstico. En este sentido, las características con valores de $r$ entre 0.3 y 0.5 presentan una correlación moderada; aquellas entre 0.5 y 0.7, una correlación fuerte; las que se encuentran entre 0.7 y 0.9, una correlación muy fuerte; y, finalmente, las comprendidas entre 0.9 y 1.0, una correlación casi perfecta.

En los resultados no se observaron características con correlación casi perfecta; sin embargo, sí se identificaron algunas con correlación alta: la desviación estándar, la media y el valor máximo del filo *Firmicutes*, así como la varianza del filo *Lentisphaerae*. Por otro lado, las características con correlación muy alta con el diagnóstico fueron la desviación estándar, el valor máximo y la media del filo *Lentisphaerae*. Por lo tanto, estas características se consideran relevantes y merecen ser destacadas.

Dentro de las mejores características el filo que tiene mayor peso en el análisis es el *Lentisphaerae* ya que los datos de media, valor máximo, desviación estándar, varianza son los que mayor peso tienen para el análisis, incluso los valores de skewness y kurtosis están dentro de las 19 características. El segundo filo de bacterias más relevante fue el *Firmicutes*

ya que los valores de media, varianza y valor máximo son los siguientes mejores después de las características del filo *Lentisphaerae*. El tercer filo de bacterias más importante fue el *Verrucomicrobia* debido a que los datos de media, desviación estándar y valor máximo son los siguientes en el peso. Después de los datos de *Verrucomicrobia* siguen los datos de desviación estándar general y de media general, siendo las métricas de toda la base de datos. El cuarto filo de bacterias que más peso tiene es el de la proteobacteria, pero solo las métricas de *kurtosis* y *skewness* tienen un peso dentro del análisis, los datos de media, varianza, desviación estándar y valor máximo de las proteobacterias no tienen un peso dentro del análisis. Los últimos datos de peso dentro del análisis son las de 2 índices de diversidad de los cinco propuestos, los cuales son el índice de Shannon y el de Riqueza de especies.

## 4.2 Visualización de los datos

A la hora de aplicar el método de LDA para la visualización de los datos sintéticos se obtuvo la gráfica de la Figura 16.

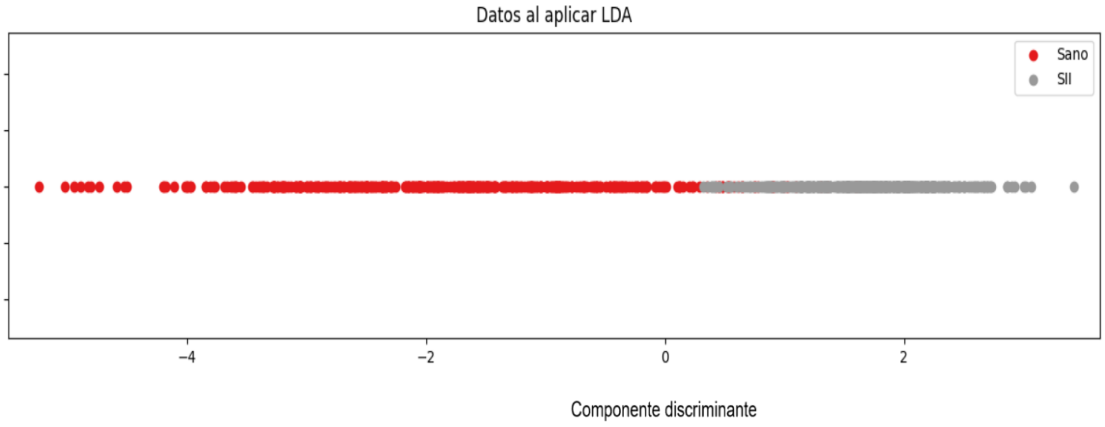



*Figura 16. Datos sintéticos en 2D al aplicar LDA*

Al aplicar el LDA para visualizar los datos reales se obtuvo la gráfica de la Figura 17.

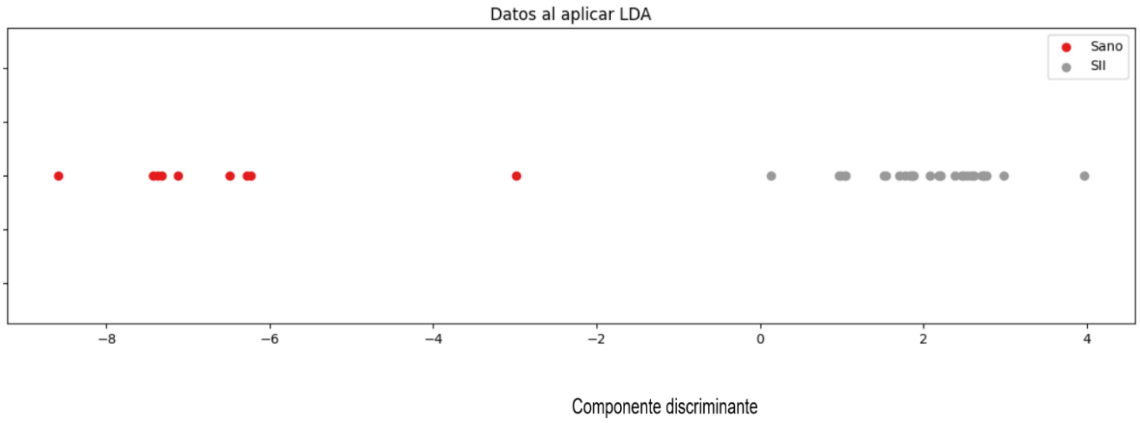



*Figura 17. Datos reales en 2D al aplicar LDA*

Al contar únicamente con dos clases, “sano” y “SII”, el número máximo de componentes discriminantes generados por el LDA es uno ($n\_clases - 1 = 1$), lo que implica que los datos se proyectan sobre una sola dimensión, quedando alineados en un eje que maximiza la separación entre ambas categorías. Como se puede apreciar, los datos se separan en 2 grupos. Aunque en el caso de los datos reales la separación es mucho mayor y en los datos sintéticos algunos se solapan. A la hora de aplicar el método de PCA para la visualización de los datos sintéticos se obtuvo la gráfica de la Figura 18.

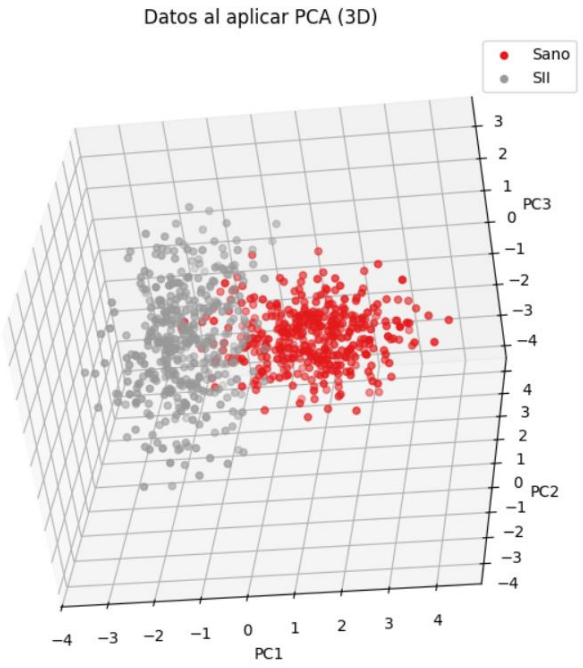



*Figura 18. Datos sintéticos en 3D al aplicar PCA*

Al aplicar el PCA para visualizar los datos reales se obtuvo la gráfica de la Figura 19.

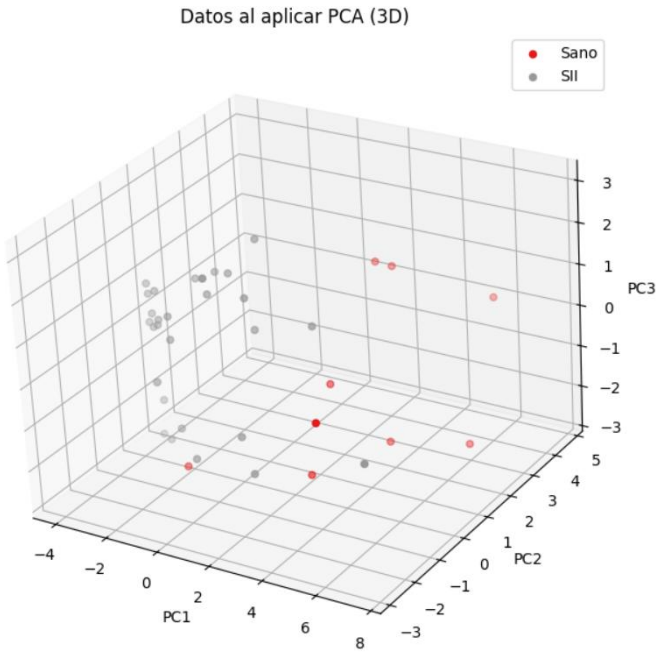



*Figura 19. Datos reales en 3D al aplicar PCA*

A la hora de usar PCA los datos sintéticos se agrupan de una manera más notoria. Mientras que los datos reales tienen una división menos definida, pero sigue siendo apreciable. Con esto se justifica la generación de datos sintéticos ya que usar pocos datos para el entrenamiento de las técnicas de IA y ML podría ser poco efectivo y llevar a un diagnóstico incorrecto del síndrome.

## 4.3 Resultados de los diagnósticos de prueba

Los resultados que se obtienen a la hora de aplicar los métodos de ML para el diagnóstico fueron los que se aprecian en la Tabla 15.

*Tabla 15. Resultados de las matrices de confusión*

| | **Verdadero Positivo** | **Falso Positivo** | **Verdadero Negativo** | **Falso Negativo** |
|---|---|---|---|---|
| *ANN-MLP* | 7 | 2 | 30 | 0 |
| *KNN* | 8 | 1 | 28 | 2 |
| *SVM* | 7 | 2 | 30 | 0 |
| *RF* | 6 | 3 | 30 | 0 |

En el campo de la medicina, los verdaderos positivos tienden a tener un gran peso a la hora de dar un diagnóstico, ya que permiten que una persona sepa si tiene una patología, por lo cual se tienden a evitar los falsos negativos porque una persona podría descartar definitivamente el tener una enfermedad que realmente sí tiene. Por otra parte, los falsos positivos tienden a ser más aceptables ya que hay estudios que pueden ayudar a corroborar los resultados, aunque lo preferible es evitarlos.

Por ello, de la tabla podemos observar en los resultados que los mejores métodos son las técnicas ANN-MLP y las SVM, ya que tienen un

número alto de verdaderos positivos sin ningún falso negativo. Mientras que el caso de las kNN tiene el mayor número de verdaderos positivos, pero también tiene el mayor número de falso negativos, por lo que no puede ser considerado como el mejor. Por otra parte, el método de RF no tiene falsos negativos, pero tienen el menor número de verdaderos positivos. En resumen, viendo estos resultados las técnicas ANN-MLP y las SVM son las mejores quedando en primer y segundo lugar, respectivamente. Por su parte, es más complicado decidir que método, kNN o RF, está en el tercer y cuarto lugar, respectivamente, por lo que es necesario usar los datos de las métricas de desempeño para decidirlo.

Los resultados de los comportamientos de los métodos de ML se aprecian en el gráfico de la Figura 20. Donde se pueden apreciar las 4 métricas de desempeño mostrados como gráfico de barras con sus respectivos valores. Estas mismas cantidades se presentan de forma resumida en la Tabla 16 para propósitos de comparación.

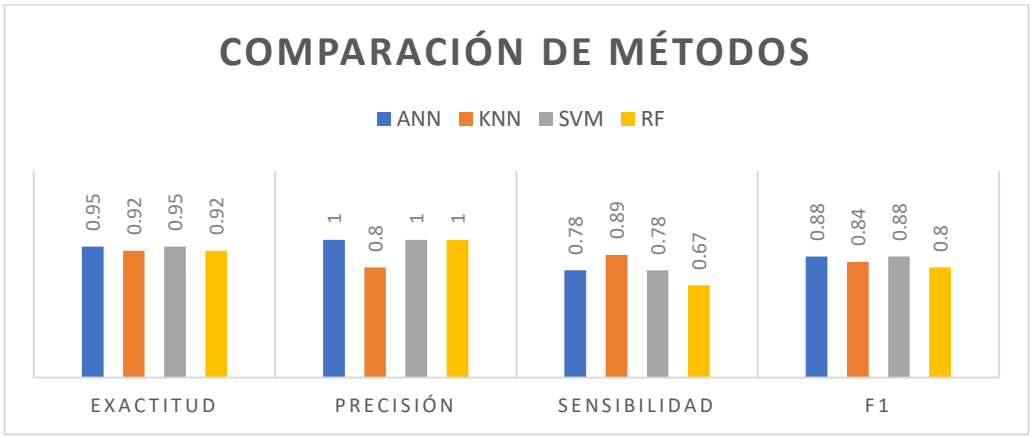



*Figura 20. Gráfico de comparación de métricas de desempeño de los cuatro métodos de ML.*

*Tabla 16. Métricas de desempeño de los métodos de ML*

| | **Exactitud** | **Precisión** | **Sensibilidad** | **F1** |
|---|---|---|---|---|
| *ANN-MLP* | 0.95 | 1 | 0.78 | 0.88 |
| *KNN* | 0.92 | 0.8 | 0.89 | 0.84 |

| | | | | |
|---|---|---|---|---|
| *SVM* | 0.95 | 1 | 0.78 | 0.88 |
| *RF* | 0.92 | 1 | 0.67 | 0.8 |

Como se aprecia en las gráficas, todos los métodos tienen métricas de desempeño superiores al 0.5, y la mayoría tienen métricas cercanas o superiores al 0.8, por lo que se infiere que los métodos de ML dan buenos resultados. No obstante, existirán desempeños superiores entre las propias técnicas y esto se analizará a continuación. Valores de exactitud superiores a 0.9 se consideran excelentes, en especial en aplicaciones médicas, por lo que todos los métodos tienen un excelente resultado en cuanto a una característica de exactitud se refiere (Buhl. 2023). Mientras que para cuestiones medicas los valores de sensibilidad y especificidad se consideran aceptables cuando su valor es muy cercano a superior al 80% (Sociedad Española de Farmacia Hospitalaria, s.f.). Así que, se puede afirmar que los 4 métodos tienen una gran probabilidad de dar un diagnóstico correcto cuando una persona está sana o cuando está enferma. Valores de precisión superiores a 0.9 se consideran excelentes, mientras que valores de precisión entre 0.8 y 0.9 se consideran buenos, por lo que los métodos de ANN-MLP, SVM y RF tienen una precisión excelente, es decir, que cuando estos métodos predicen que una persona está enferma suelen tener razón, evitando los falsos positivos. Mientras tanto, el resultado de kNN tiene un valor entre 0.8 y 0.9 por lo que se considera bueno, es decir que puede tener cierto margen de error al diagnosticar que una persona tiene SII. Los valores de sensibilidad de los métodos de ANN-MLP, kNN y SVM están en el rango de 0.8 y 0.9, implicando que estas tres técnicas tienen una adecuada pero no alta sensibilidad, es decir, una buena capacidad de discriminar resultados (o realizar el diagnóstico del SII) ante cambios incipientes de los datos de los usuarios. Mientras que el método de RF tiene un valor entre 0.5 y 0.8, indicando que este método es el que más baja sensibilidad tiene ante variaciones de los datos procesados, podría implicar que para un mejor

resultado esta técnica requiere que los datos muestren una variación más marcada en los indicadores biológicos o estadísticos. Finalmente, los valores de F1 de todos los métodos tienen valores entre 0.8 y 0.9 por lo que se considera que tienen un buen resultado de F1, con estos valoresse puede decir que estos métodos tienen un buen equilibrio entre detectar a los pacientes que verdaderamente están enfermos y evitar falsos positivos en el diagnóstico.

Los métodos que mejor resultado tuvieron fueron el ANN-MLP y el SVM teniendo resultados excelentes en las métricas de exactitud y precisión y resultados buenos en las métricas de sensibilidad y F1. El método de kNN tiene el tercer lugar de mejor método al tener un resultado excelente en las métricas de exactitud y buenos resultados en las métricas de precisión, sensibilidad y F1. El método que podría ser considerado como el peor es el de RF al tener resultados excelentes exactitud y precisión, buen resultado en F1, pero un resultado aceptable de sensibilidad.

## 4.4 Herramienta de diagnóstico

La herramienta que se desarrolló para ayudar en el diagnóstico del SII cuenta con 3 pestañas como se plasma en el diagrama de flujo de la sección 3.7. La primera pestaña de la herramienta es la de inicio, mostrada en la Figura 21, la cual es para describir el origen y la finalidad del proyecto y la naturaleza de este.

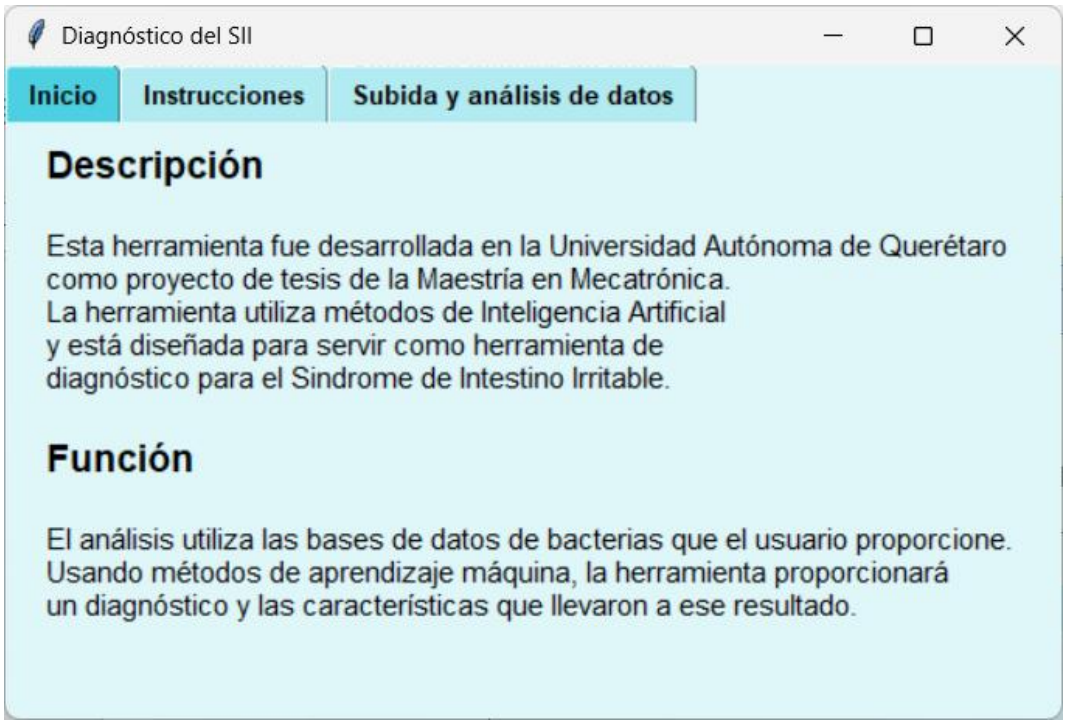
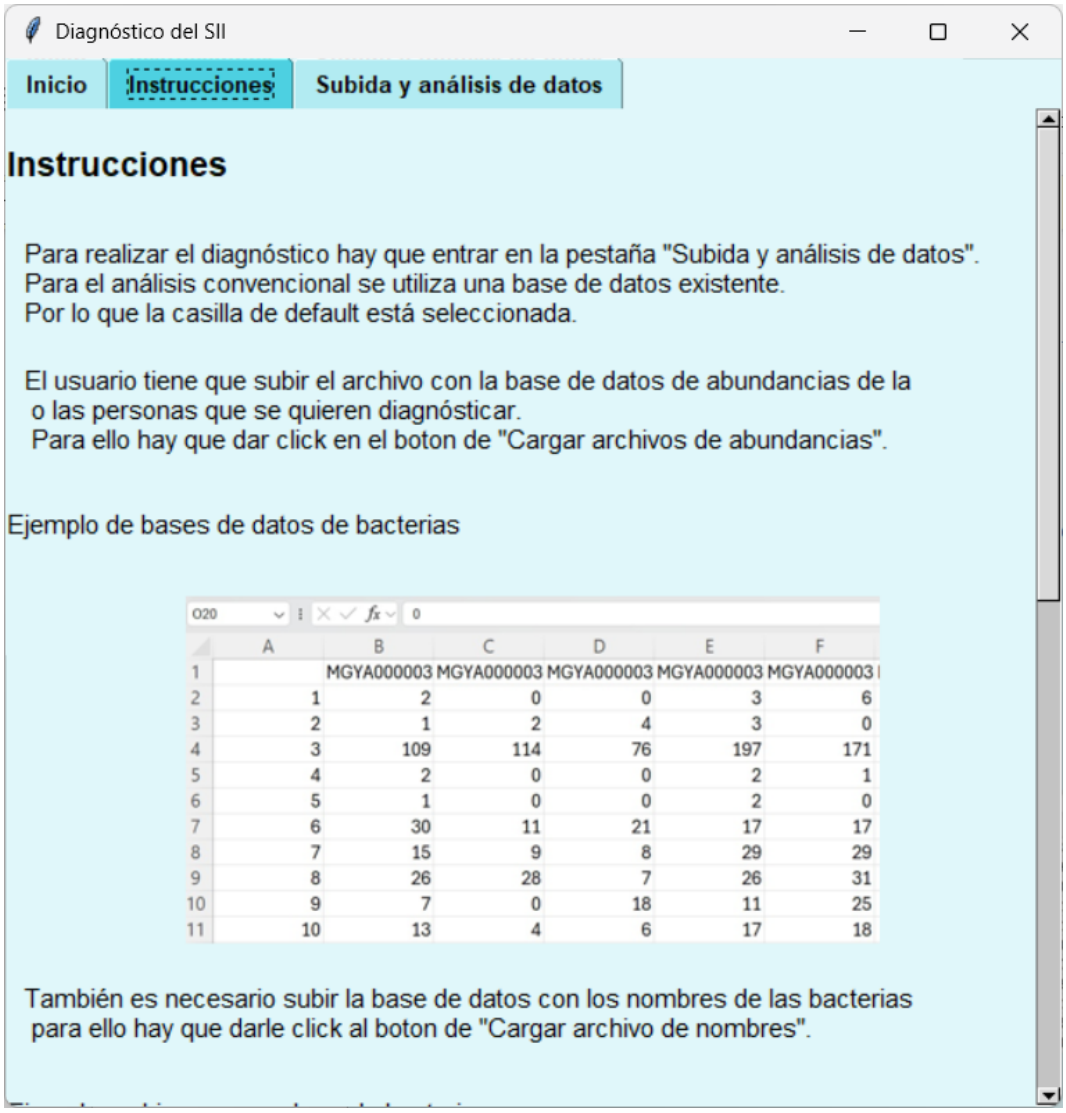



*Figura 21. Página de Inicio de la herramienta de diagnóstico*

La segunda pestaña de la interfaz de usuario es la sección de instrucciones que sirve de guía para el usuario. En esta sección se describe como subir y procesar los datos y como utilizar las funciones que están presentes en la interfaz. Además, muestra un ejemplo de cómo deben estar estructurados los datos que subes a la herramienta de diagnóstico para su correcto funcionamiento. En las imágenes 22 y 23 se muestra la sección de Instrucciones de la herramienta de diagnóstico.




| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | | MGYA000003 | MGYA000003 | MGYA000003 | MGYA000003 | MGYA000003 |
| 2 | 1 | 2 | 0 | 0 | 3 | 6 |
| 3 | 2 | 1 | 2 | 4 | 3 | 0 |
| 4 | 3 | 109 | 114 | 76 | 197 | 171 |
| 5 | 4 | 2 | 0 | 0 | 2 | 1 |
| 6 | 5 | 1 | 0 | 0 | 2 | 0 |
| 7 | 6 | 30 | 11 | 21 | 17 | 17 |
| 8 | 7 | 15 | 9 | 8 | 29 | 29 |
| 9 | 8 | 26 | 28 | 7 | 26 | 31 |
| 10 | 9 | 7 | 0 | 18 | 11 | 25 |
| 11 | 10 | 13 | 4 | 6 | 17 | 18 |

*Figura 22. Sección de instrucciones de la herramienta de diagnóstico - A*

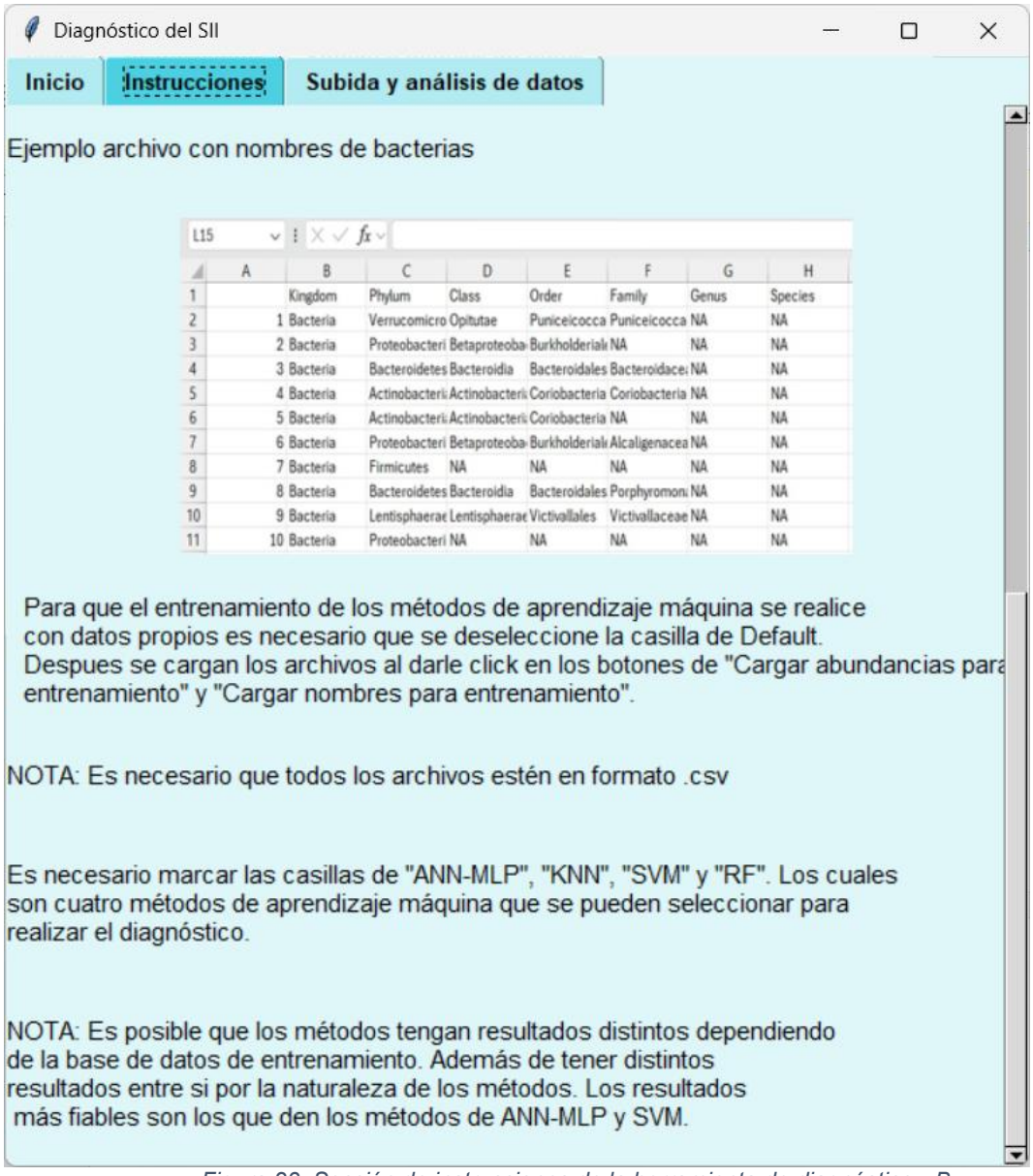



*Figura 23. Sección de instrucciones de la herramienta de diagnóstico - B*

La tercera pestaña es la más importante, ya que consiste en la interfaz de usuario donde se pueden subir los datos de un paciente y obtener un diagnóstico indicando si tiene o no el SII. Por default, está marcada la opción de usar los mismos datos de entrenamiento que se emplearon en el presente trabajo, el usuario solo necesita subir los datos del individuo para hacer el diagnóstico al dar click en los botones de “Cargar archivo de abundancias” y “Cargar archivo de nombres”. Sin embargo, el usuario puede optar por utilizar

los datos propios para entrenar los algoritmos al desmarcar la opción en la casilla de “Default” y dar click en los botones de “Cargar abundancias para entrenamiento” y “Cargar nombres para entrenamiento”. También se cuenta con la opción de elegir entre los dos métodos para la visualización de los datos, las casillas de LDA y PCA, que permiten visualizar los datos en 2D y 3D respectivamente, aunque si no se selecciona ninguno simplemente no se visualizaran los datos. La posibilidad de aplicar una previsualización de los datos usando ya sea LDA o PCA, podría servir como un primer indicativo de si los datos tienen o no características discriminantes para diagnosticar el SII, faltaría confirmar el diagnóstico con exactitud y precisión al aplicar la técnica de ML. A la derecha de las casillas de LDA y PCA se encuentran las opciones de selección de los métodos de ML empleados en este trabajo. Cuando se utiliza la base de datos por defecto, se recomienda al usuario seleccionar los métodos con mejor desempeño, ANN-MLP y SVM, aunque también es posible elegir cualquiera de los 4 métodos disponibles, de manera individual o combinada. Es importante señalar que, si no se selecciona ninguno, no se generará un resultado de diagnóstico. En caso de trabajar con bases de datos distintas a la predeterminada, se sugiere aplicar los 4 métodos para obtener un análisis más completo. En el cuadro de texto de la parte inferior de la interfaz se muestran las características que más peso tuvieron para la clasificación de la base de datos de entrenamiento al darle click en el botón de análisis. También, en el mismo cuadro de texto y al darle click al botón de análisis, se muestran los resultados de diagnóstico que proporcionan los métodos de ML marcados en las casillas siendo el valor de 1 una persona sana y un valor de 2 para una persona con SII. En las imágenes 24 y 25 se muestra la sección de análisis de datos de la herramienta de diagnóstico con todos los componentes ya mencionados.

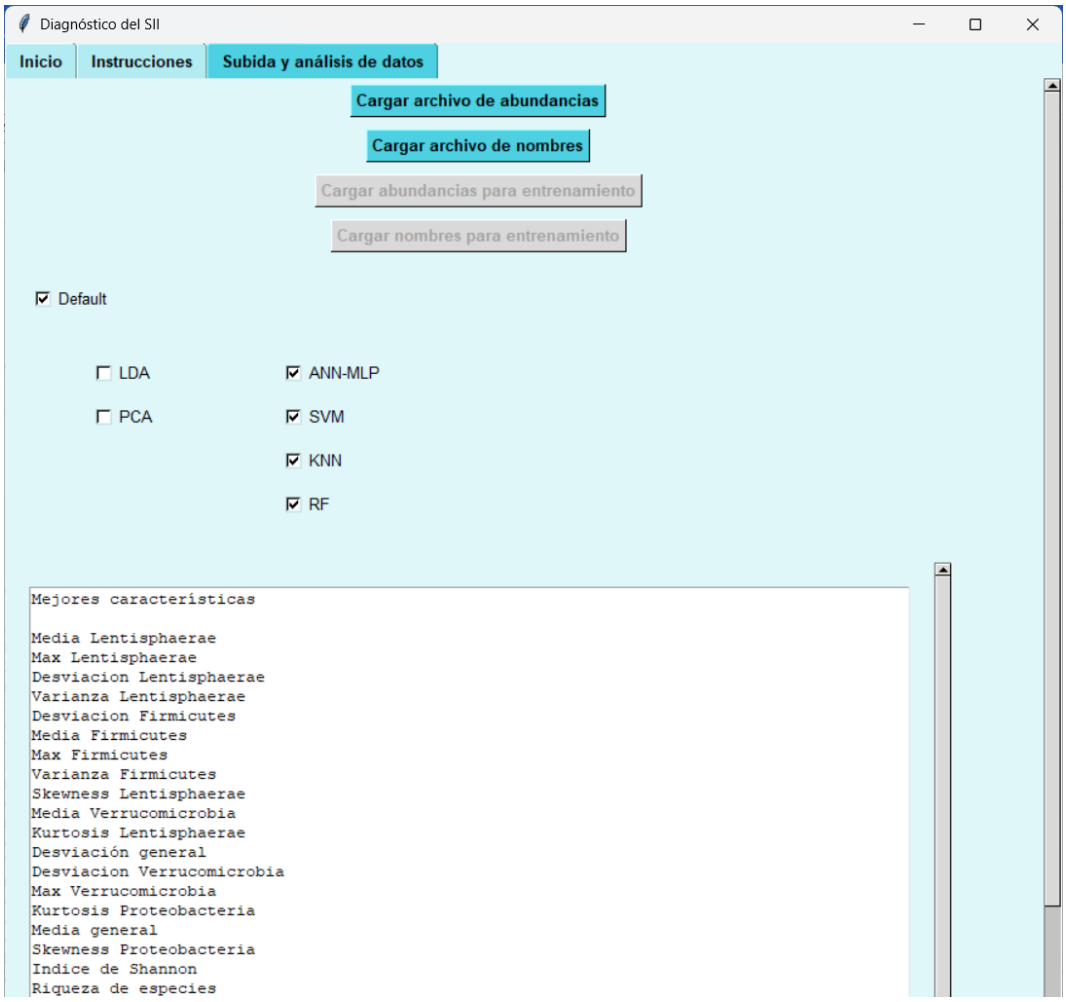



*Figura 24. Página de análisis de datos de la herramienta de diagnóstico - A*

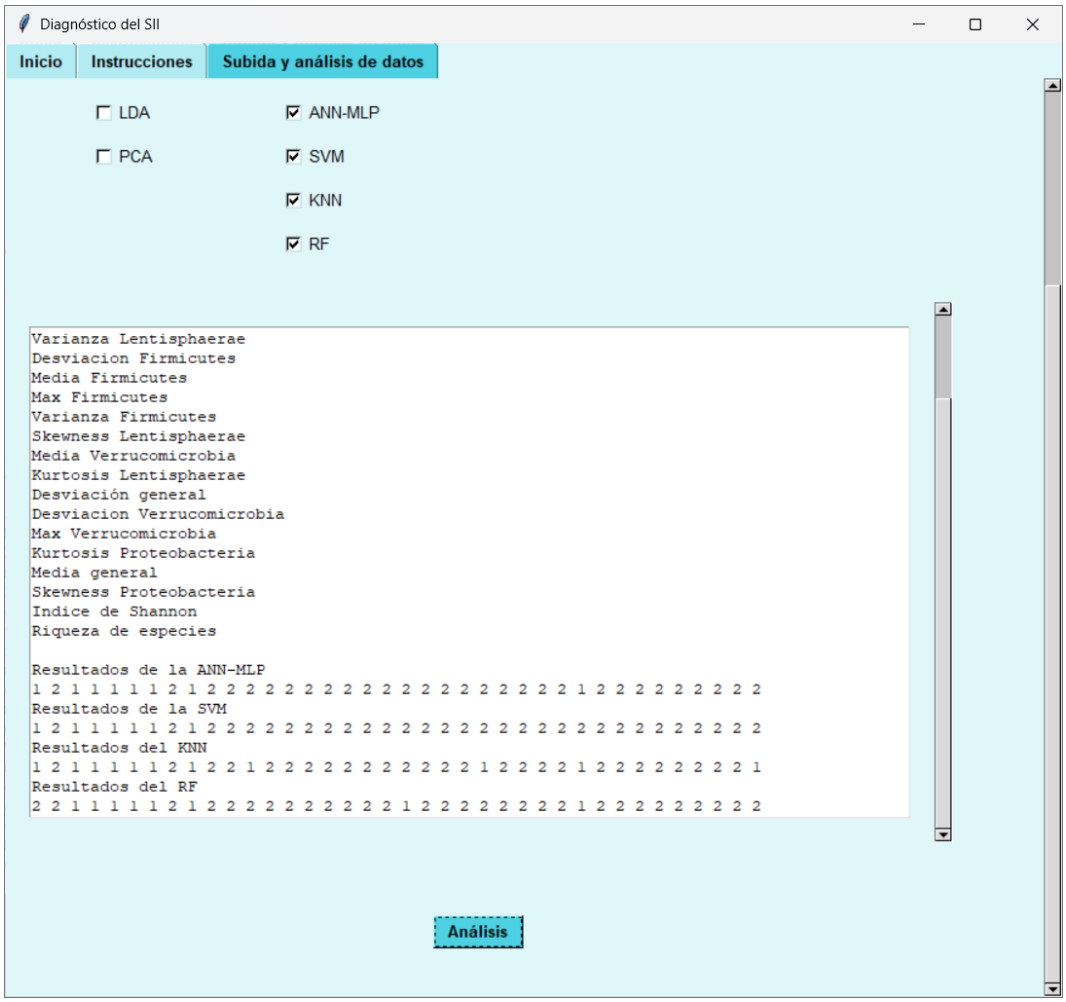



*Figura 25. Página de análisis de datos de la herramienta de diagnóstico – B*

# CAPITULO 5: CONCLUSIONES

Todos los métodos obtuvieron una exactitud superior al 92%, por lo que las métricas elegidas para el entrenamiento son acertadas y los métodos tienen excelentes valores de exactitud. Aunque no es posible comparar estos resultados con los métodos actuales, el criterio de roma III o el criterio de roma IV, debido a que no hay estudios en los que se pruebe su exactitud, solamente los valores de precisión, sensibilidad y F1.

Los métodos que mejores resultados en general obtuvieron para diagnosticar el SII fueron las técnicas de ANN-MLP y SVM. La precisión de ambas técnicas es 100%, lo cual supera al 82.4% del criterio de roma IV, el método más utilizado de la actualidad. Es posible que sea porque, como se ven en las representaciones 2D y 3D al usar LDA y PCA, los datos tienden a agruparse y tienen una separación visible. El peor método de ML para diagnosticar el síndrome fue RF, tal vez porque el método toma los indicadores calculados (estadísticos y de diversidad) al azar para crear los árboles de decisión, mientras que otras características que pudieran contener información discriminante valiosa no entran en el análisis. Como todas las características utilizadas mostraron ser relevantes, el dejar fuera a algunas de ellas para el algoritmo puede haber afectado negativamente su capacidad de clasificación. La técnica kNN no proporcionó buenos resultados, esto se puede atribuir a que el número de características que utiliza es demasiado alto, por lo tanto, se genera una carga de información redundante que incapacita a la técnica de dar resultados confiables o provoca sobreajuste de resultados.

Por otra parte, la sensibilidad de las técnicas ANN-MLP y SVM es del 78%, lo cual no supera el 82.9% del criterio de roma IV, el método más utilizado, pero se acerca bastante. El método que mejor sensibilidad obtuvo

fue la kNN con un 89%, el cual es superior a la sensibilidad del mejor método actual que es de 82.9%. Aunque su precisión solo es del 80%. El método de RF tuvo una precisión del 100% pero su sensibilidad fue la más baja de los 3 métodos con un 67%.

Los métodos de ANN-MLP y SVM son los más altos en el score F1 con valores del 88% mientras que el kNN y el RF tenían valores de 84% y 80% respectivamente. Por lo que las ANN-MLP serían los mejores métodos para este caso y con esta base de datos. El mejor método actual y el más usado, el criterio de roma IV, tiene un F1 de 82.7% mientras que 3 de los métodos basados en ML usados en el presente trabajo los superan y 2 de ellos llegan al 88%.

Los métodos actuales tienen un proceso de diagnóstico de hasta 3 meses o incluso más. Por otro lado, el método utilizado solo depende de los tiempos del laboratorio para sacar las muestras necesarias.

El que las técnicas de ANN-MLP y SVM tengan mejor resultado puede deberse a que estos métodos pueden usar pocas muestras para dividir los grupos, mientras que el RF necesita una mayor cantidad de datos. El kNN puede no dar tan buenos resultados debido a que el número de características es demasiado alto.

Los índices de diversidad que eran significantes para este análisis son el índice de Shannon y la riqueza de especies por lo que estos índices podrían ser utilizados en futuros análisis de este tipo.

Los filos de bacterias cuyas características tenían un mayor peso en el diagnóstico fueron las *Lentisphaerae*, *Firmicutes* y *Verrucomicota*. Por lo

que estos filos podrían estudiarse más a fondo en trabajos futuros para encontrar posibles tratamientos para el SII.

El que la media general sea una de las 19 características que más peso dan, indica que la reducción de bacterias en general es un indicador de que una persona tiene SII.

Los tres filos con mayor peso tienen todas sus métricas, o casi todas, dentro de las 19 mejores características, por lo que el filo en si tiene un peso dentro del análisis, pero en el caso del filo de *Protobacteria* solo la kurtosis y el skewness. Por lo que si una o pocas de las especies de este filo están muy altas o bajas respecto a las otras es un indicador de que la persona tiene SII.

Los resultados obtenidos por la herramienta son bastante buenos considerando que se emplearon datos sintéticos, alcanzando valores de sensibilidad y precisión cercanos o superiores al 80%. Sin embargo, el estudio de Su et al. (2022), que presenta los mejores resultados en la detección del SII mediante IA, reportó valores aún mayores, probablemente debido a que utilizó una base de datos amplia y compuesta por datos reales.

# CAPITULO 6: PROSPECTIVAS

Una base de datos con una mayor cantidad de muestras y que tenga un grupo de control de una buena dimensión podría mejorar el rendimiento de la herramienta. Incluso se podría probar el juntar bases de datos obtenidas por diferentes laboratorios y comprobar la eficacia de la herramienta al tener abundancias con una normalización distinta y un número y tipo de especies diferentes.

En el presente trabajo solo se utilizaron los métodos de LDA y PCA con fines de visualización, pero en trabajos similares podrían utilizarse estos métodos en conjunto con los métodos de ML. Es decir, aplicar estos métodos de reducción de la dimensionalidad a los datos y después aplicar los métodos de diagnóstico sobre los datos transformados, ya que las técnicas LDA y PCA indican propiedades discriminatorias de forma gráfica. Tal vez se podría conseguir mejores resultados con algunos métodos.

Aunque se utilizaron 4 métodos de ML en este trabajo por ser los más utilizados en el ámbito de la medicina, también se podrían aplicar otros métodos de ML o combinaciones de estos y probar su eficacia, en especial si se manejan bases de datos más grandes.

Hay enfermedades similares al SII como puede ser la enfermedad de Crohn, por lo que se pueden confundir. Sería interesante probar la robustes y la capacidad de la herramienta creada con estas enfermedades para comprobar si puede diferenciar al SII de otras patologías.

La herramienta podría ser usada para trabajos futuros como base de otros trabajos donde se eliminen o se añadan otras características, como otros filos de bacterias u otros índices de diversidad, e incluso ser mejorada

estéticamente para que sea más atractiva para algún centro médico y se implemente como una página web de libre acceso.

Los filos de bacterias que se encontraron en este trabajo podrían servir para que trabajos futuros las estudien y encuentren un tratamiento efectivo. Incluso se podrían analizar las mejores características para mejorar la metodología presente al eliminar las características y filos prescindibles, reduciendo el costo computacional de la herramienta.

# REFERENCIAS BIBLIOGRÁFICAS


*Norman, G. R., Monteiro, S. D., Sherbino, J., Ilgen, J. S., Schmidt, H. G., & Mamede, S. (2017). The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. Academic Medicine, 92(1), 23-30.*

*Bassotti, G. (2022). Irritable Bowel Syndrome: A Multifaceted World Still to Discover. Journal of Clinical Medicine, 11(14), 4103.*

*Bhavsar, K. A., Abugabah, A., Singla, J., AlZubi, A. A., & Bashir, A. K. (2021). A comprehensive review on medical diagnosis using machine learning. Computers, Materials and Continua, 67(2), 1997.*

*Mazo, C. M. G. (2025). Aplicación del Modelo Media Varianza con Machine Learning para Optimización de Portafolios de Inversión. European Public & Social Innovation Review, 10, 1-20.*

*García-Basurto, A., Saucedo-Dorantes, J. J., Pérez-Cruz, Á., & Osornio-Ríos, R. A. (2021). Análisis de falla de encendido en motores de combustión utilizando señales de vibración basado en el cálculo y reducción de indicadores estadísticos. Científica, 25(1), 01-11.*

Domingo, J. J. S. Síndrome del intestino irritable. *Medicina Clínica (English Edition)*. 158(2), 76-81. 2021.

Black CJ, Craig O, Gracie DJ, Ford AC. Comparison of the Rome IV criteria with the Rome III criteria for the diagnosis of irritable bowel syndrome in secondary care. Gut. 2021 Jun;70(6):1110-1116. 2020 Sep 24.

Alvarado, J., Otero, W., Jaramillo Santos, M. A., Roa, P. A., Puentes, G. A., Jiménez, A. M., ... & Sabbagh, L. (2015). Guía de práctica clínica para el diagnóstico y tratamiento del síndrome de intestino irritable en población adulta. *Revista colombiana de Gastroenterología*, *30*, 43-56.

Shaikh, S.D.; Sun, N.; Canakis, A.; Park, W.Y.; Weber, H.C. Irritable bowel syndrome and the gut microbiome: A comprehensive review. *Journal of Clinical Medicine*. 12(7). 2023.

Otero W. y Gómez M. Síndrome de Intestino Irritable: diagnóstico y tratamiento farmacológica Revisión concisa. *Revista de gastroenterología del Perú*. 25. 2005.

PASTOR, G. P., OTERO, B. M., PRATS G. & MIRELIS B. Diversidad bacteriana. Principales bacterias en patología humana. 2002.

Fukui, Hirokazu, et al. "Usefulness of machine learning-based gut microbiome analysis for identifying patients with irritable bowels syndrome." *Journal of clinical medicine* 9.8. 2020.

Su, Qi, et al. "Faecal microbiome-based machine learning for multi-class disease diagnosis." *Nature Communications* 13.1. 2022.

Ghaffari P., Shoaie S., Nielsen L. Irritable bowel syndrome and microbiome; Switching from conventional diagnosis and therapies to personalized intervetions. *J Transl Med.* 20, 173. 2022.

Remes J., Gómez O., Nogueira J., Carmona R., Pérez J., López A., Sanjurjo J., Noble A., Chávez J. y González M. Tratamiento farmacológico del síndrome de intestino irritable: revisión técnica. *Rev Gastroenterol Mex*. 75(1). 2010.

Chong P., Chin V., Looi C., Wong W., Madhavan P. y Yong V. The Microbiome and Irritable Bowel Syndrome - A review on the Pathophysiology, current research and future therapy. *Front. Microbiol.* 10, 1136. 2019.

Arce W. Disbiosis intestinal: alteración de la relación mutualista entre la microbiota y sistema inmune. *Acta académica*. 67(noviembre), 171-182. 2020.

Álvarez J., Fernández J., Guarner F., Gueimonde M., Rodrígez J., De Pipaon M. y Sanz Y. Microbiota intestinal y salud. *Gastroenterología y hepatología*. 44(7), 519-535. 2021.

Sakamoto K., Arias J. y Moreno F. Relación entre la colonización de la microbiota intestinal y el desarrollo de patologías inflamatorias intestinales. *Salutem Scientia Spiritus*, *8*(4), 56–63. 2022.

Icaza M. Microbiota intestinal en la salud y la enfermedad. Revista de gastroenterología de México. 78(4), 240-248. 2014.

Machado M., Mora G. y Peña S. Implicación de la disbiosis intestinal en la obesidad. *MQRInvestigar*, *7*(2), 1215–1240. 2023.

Fontané L., Benaiges D., Goday A., Llauradó G. y Botet J. Influencia de la microbiota y de los probióticos en la obesidad. *Clínica e investigación en arteriosclerosis*. 30(6), 271-279. 2018.

Chan Y. Estaki M. y Gibson D. Consecuencias clínicas de la disbiosis inducida por la dieta. *Annales Nestlé*. 63(suppl), 28-40. 2013.

Huang W., Wu J., Mao Y., Zhu S., Huang G., Petritis B. y Huang R. Developing a periodontal disease antibody array for the prediction of severe periodontal disease using machine learning classifiers. *Journal of periodontology*. 91(2), 232-243. 2019.

Zia H. Análisis automático de imágenes de frotis de sangre periférica para diagnóstico de Leucemia [Trabajo de fin de grado, Universidad de Navarra]. 2021.

Aguirre F., Carballo L., González X. y Gigirey V. Inteligencia Artificial aplicada a la Figura médica. *Revista de Figuraología*. 24(2), 9-20. 2021.

Zand A., Stokes Z., Sharma A., Van Deen W. y Hommes D. Artificial Intelligence for Inflammatory Bowel Diseases (IBD); Accurately

Predicting Adverse Outcomes Using Machine Learning. *Dig Dis Sci.* 67, 4874–4885. 2022.

Cruz J., Hernández P., Dueñas N. y Salvato A. Importancia del método clínico. *Revista Cubana de Salud Pública*. 38, 422-437. 2012.

Díaz J., Gallego B. y León A. El diagnóstico médico: Bases y procedimientos. *Revista cubana de medicina general integral*. 22(1). 2006.

Lorenzano C. El diagnóstico médico. Subjetividad y procesos cognitivos. 2006.

Borja, R., Monleón, A., & Rodellar, J. Estandarización de métricas de rendimiento para clasificadores Machine y Deep Learning. *Revista Ibérica de Sistemas e Tecnologias de Informação*. (E30), 184-196. 2020.

Serrano, Francisco Emiliano Aguayo, et al. "Desarrollo de un proceso de autenticación facial en un sistema android utilizando el algoritmo lda (análisis de discriminación lineal)." *Pistas Educativas* 39.128. 2018.

Moreno X. y Vialva A. Generalidades de la microbiota intestinal. *Caracas: Sociedad Venezolana de Bioanalistas Especialistas*. 27-34. 2019.

Moreno X. Disbiosis en la microbiota intestinal. *Revista GEN*. 76(1), 17-23. 2022.

Random Forest Algorithm Overview (H. A. Salman, A. Kalakech, & A. Steiti , Trans.). (2024). *Babylonian Journal of Machine Learning*, *2024*, 69-79.

Suárez, E. J. C. (2014). Tutorial sobre máquinas de vectores soporte (sVM). *Tutorial sobre Máquinas de Vectores Soporte (SVM).*

Serrano, F. E. A., Ortega, J. C. P., Fernández, M. A. A., & Hurtado, E. G. (2020). Reducción y extracción de características faciales en imágenes utilizando análisis discriminante lineal (LDA) y análisis del

componente principal (PCA). *Perspectivas de la Ciencia y la Tecnología*, *3*(6), 64-76.

González J. Midiendo la diversidad biológica: más allá del índice de Shannon. *Acta zoológica lilloana*, 3-14. 2012.

Otero, J., Sánchez, A. H., & Moral, E. M. (2005). Análisis de la varianza (ANOVA). *DOCPLAYER. Obtenido de https://docplayer. es/10487925-Analisis-de-la-varianza-anova-jose-vicens-otero-ainhoa-herrarte-sanchez-eva-medina-moral. html*.

Matich D. Redes Neuronales: Conceptos Básicos y Aplicaciones. *Acta zoológica lilloana*, 3-14. 2001.

Borja R., Monleón A. y Rodellar J. Estandarización de métricas de rendimiento para clasificadores Machine y Deep Learning. *Revista Ibérica de Sistemas e Tecnologias de Informação*, (E30), 184-196. 2020.

Stoppani, C. L. Efecto de Lactobacillus Salivarius sobre la microbiota intestinal, el estado sanitario y el desempeño productivo de cerdos en etapa de cría [Tesis doctoral, Universidad de Buenos Aires]. 2013.

Golondrino Rernández, L. G. Clasificación taxonómica de bacterias usando machine learning [Trabajo de fin de grado, Universidad de los Andes]. 2020.

Lázaro V. Redes multicapa como herramienta de análisis de la microbiota [Trabajo de obtención de grado, Universidad Autónoma de Querétaro]. 2022.

Patón González, V. Uso de inteligencia artificial para prevenir la formación de fístulas colónicas y su relación con la microbiota intestinal [Trabajo Fin de Grado, E.T.S. de Ingeniería Agronómica, Alimentaria y de Biosistemas (UPM)]. 2021.

Velasco M., Piles M., Viñas M., Rafel O., Gónzalez O., Guivernau O. y Sánchez J. Determinismo genético de la microbiota intestinal del conejo. *XIX Reunión Nacional de Mejora Genética Animal.* 2018.

Rajotte, J. F., Bergen, R., Buckeridge, D. L., El Emam, K., Ng, R., & Strome, E. (2022). Synthetic data as an enabler for machine learning applications in medicine. *Iscience*, *25*(11).

Buhl, N. *F1 Score in Machine Learning Explained*. Encord. 2023.

Sociedad Española de Farmacia Hospitalaria. *Validez de la prueba diagnóstica: Sensibilidad y especificidad.* s.f.