UNIVERSIDAD AUTÓNOMA DE ZACATECAS

“Francisco García Salinas”

Uso de Metabolitos para el Desarrollo de Modelo de Aprendizaje Automático para el Diagnóstico y Predicción de Enfermedad Grave del COVID-19

Tesis para obtener el grado de:

Maestro en Ciencias del Procesamiento de la Información

Presenta

Hugo Alexis Torres Pasillas

Director:

Dr. José María Celaya Padilla

Co-Directores:

Dr. Yamilé López Hernández

Carlos Eric Galván Tejada

Asesores:

Dra. Alejandra García Hernández

Dr. Pedro Daniel Alaniz Lumbreras

Zacatecas, Zac., junio de 2024

Zacatecas, Zac., 22 de mayo de 2024.

Lic. Hugo Alexis Torres Pasillas
Estudiante de la MCPI
PRESENTE

Dr. Huizilopoztli Luna García
Responsable de la MCPI

Nos es grato comunicarle que después de haber sometido a revisión académica la propuesta de Tesis Titulada **"Uso de metabolitos para el desarrollo de modelo de aprendizaje automático para el diagnóstico y predicción de enfermedad grave del COVID-19"**, presentada por el estudiante **Lic. Hugo Alexis Torres Pasillas** y habiendo efectuado todas las correcciones indicadas por este Comité Tutorial, se **AUTORIZA** el documento de tesis para su impresión.

Sin más por el momento reciba un cordial saludo.

COMITÉ TUTORIAL
PROCESAMIENTO Y ANÁLISIS DE DATOS

Dr. José María Celaya Padilla

Dra. Yamilé López Hernández

Dr. Carlos Eric Galván Tejada

Dra. Alejandra García Hernández

Dr. Pedro Daniel Alaniz Lumbreras

Carta de similitud núm.634/ IyP
Zacatecas, Zacatecas 23/mayo/2024

Dr. Huizilopoztli Luna García
Responsable de la MCPI – UAZ
Presente

Estimado Dr. Huizilopoztli,

Después de saludarlo, sirva el presente oficio para notificar que el documento

"Uso de metabolitos para el desarrollo de modelo de aprendizaje automático para el diagnóstico y predicción de enfermedad grave del COVID-19"
de Hugo Alexis Torres Pasillas

Fue analizado con el software Copyleaks, con la intención de detectar similitudes; el resultado en cuestión fue

6 % de similitud

De acuerdo a lo anterior, el porcentaje se considera **ACEPTABLE** de acuerdo a los estándares internacionales.

Atentamente
"Somos Arte, Ciencia y Desarrollo Cultural"

Dr. Carlos Francisco Bautista Capetillo
Coordinador de Investigación y Posgrado
Universidad Autónoma de Zacatecas

Zacatecas, Zac., 22 de mayo de 2024.

Carta Cesión de Derechos

A QUIEN CORRESPONDA

El que suscribe C. **Hugo Alexis Torres Pasillas** alumno del **Programa Maestría en Ciencias del Procesamiento de la Información** con número de matrícula **42107919**, adscrito a la Unidad Académica de Ingeniería Eléctrica de la Universidad Autónoma de Zacatecas, manifiesta que es autor intelectual del presente trabajo de Tesis bajo la dirección del Dr. José María Celaya Padilla y cede los derechos del trabajo titulado "**Uso de metabolitos para el desarrollo de modelo de aprendizaje automático para el diagnóstico y predicción de enfermedad grave del COVID-19**" a la Universidad Autónoma de Zacatecas para su difusión, con fines académicos y de investigación.

ATENTAMENTE

Hugo Alexis Torres Pasillas

AGRADECIMIENTO ESPECIAL

Al Pograma Nacional de Posgrados de Calidad (PNPC) del Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT) por su apoyo económico a través de la convocatoria "Becas Nacionales (Tradicional) 2022-2024.

Agradecimientos

Quiero expresar mis agradecimientos a todo el personal del programa de Maestría en Ciencias de la Información. Su dedicación y compromiso son invaluables para la calidad del programa y el crecimiento de los estudiantes.

A todos mis profesores, que siempre buscan compartir sus conocimientos con una alta pasión y dedicación a la enseñanza, quienes con su entusiasmo contagioso y amplia experiencia me han guiado en este camino, de quienes he aprendido muchísimas cosas de cada uno de ellos no solo en el ámbito académico, sino también como persona.

Un especial agradecimiento a mi director de tesis, el Dr. José María Celaya Padilla, quien ha sido una gran persona tanto como profesor como asesor, y quien me ayudó ampliamente a desarrollar este trabajo de investigación. Así mismo, quiero agradecer ampliamente a mis codirectoras, la Dra. Yamilé López Hernández, y el Dr. Carlos Eric Galván Tejada, así como a mis asesores, la Dra. Alejandra García Hernández y el Dr. Pedro Daniel Alaniz Lumbreras, quienes de diferentes maneras han sido parte de este trabajo.

De manera muy especial, quiero agradecer a todos mis amigos, quienes han sido compañeros incondicionales en este proceso de mi formación profesional. A lo largo de la maestría, hemos compartido momentos invaluables que han enriquecido mi vida y me han brindado un apoyo invaluable.

De manera muy especial, quiero expresar mi más profundo agradecimiento a mi amada familia, quienes han sido el pilar fundamental que me ha sostenido durante este camino de formación profesional. Su amor incondicional, su apoyo inquebrantable y su constante aliento han sido la fuerza que me ha impulsado a alcanzar mis metas y superar los obstáculos.

Dedicatoria

Dedicada a mi familia.

Resumen

El surgimiento del virus SARS-CoV 2 ha sido causante de una de las peores crisis sanitarias globales de las últimas décadas. A pesar de los distintos esfuerzos realizados por investigadores, médicos y otros sectores en todo el mundo, los resultados no han sido tan satisfactorios, evidenciando que el sistema de salud como se ha concebido hasta el momento no se encuentra completamente preparado para enfrentar este tipo de situaciones. Desde el inicio de la pandemia, uno de los principales causantes de su alta propagación alrededor del mundo fue la falta de métodos de diagnóstico eficaces, fáciles de producir y distribuir, que permitieran la detección de una gran proporción de las personas infectadas. Así mismo, se carecía de métodos para inferir el nivel de severidad de las enfermedades en los pacientes, lo que dificultó la gestión más efectiva de los recursos disponibles. Por ello, en este trabajo realizamos una investigación sobre el uso de modelos de aprendizaje automático, combinado con datos del perfil metabólico de los pacientes, para su uso como herramientas eficaces de diagnóstico y pronóstico (predicción de nivel de severidad en pacientes enfermos) de COVID-19. Los resultados muestran modelos confiables, capaces de diagnosticar con una alta precisión y sensibilidad a los pacientes. Durante el entrenamiento, el modelo alcanza una exactitud balanceada de 95.8% (precisión=98.8%, sensibilidad = 92.9% y especificidad=96.0%), mientras que en conjunto de prueba (con datos ciegos), la exactitud balanceada es de 91.7% (sensibilidad, especificidad y precisión=0.91.7%). Por otro lado, el modelo obtiene también un desempeño alto para clasificar a los pacientes enfermos que desarrollan una enfermedad grave (precisión, sensibilidad, y exactitud balanceada de 79%). Así, el modelo propuesto muestra un alto potencial como una herramienta médica alternativa a las existentes o bien como un posible enfoque en enfermedades futuras, con un desempeño similar a los logrados por dichas técnicas, y con la capacidad de realizar también una predicción del nivel de severidad de los enfermos.

Palabras clave: COVID-19, diagnóstico, pronóstico, aprendizaje automático, metabolitos, selección de características.

Abstract

The emergence of the SARS-CoV-2 virus led to one of the worst global health crises in recent decades. Although numerous efforts were made by researchers, doctors, and other sectors around the world, the results have not been as satisfactory, highlighting that our healthcare system is not fully prepared to face these types of situations. Since the onset of the pandemic, one of the main contributors to its high spread worldwide was the lack of effective diagnostic methods that are easy to produce and distribute, allowing for the detection of a large proportion of infected individuals. Additionally, there was a lack of methods to infer the severity level of illnesses in patients, which hindered the more effective management of available resources. Therefore, in this study, we evaluated the use of machine learning models combined with patients' metabolic profile data for its use as effective diagnostic and prognostic tools (predicting severity level in sick patients) for COVID-19. The results show reliable models capable of diagnosing patients with high precision and sensitivity. During training, the model achieves a balanced accuracy of 95.8% (precision=98.8%, sensitivity=92.9%, and specificity=96.0%), while in the test set (with blind data), the balanced accuracy is 91.7% (sensitivity, specificity, and precision=91.7%). On the other hand, the model also performs well in classifying sick patients who develop severe illness (precision, sensitivity, and balanced accuracy of 79%). Thus, the proposed model demonstrates high potential as a medical tool alternative to existing ones or as a possible approach for future disease diagnosis, with performance similar to that achieved by said techniques, and with the capability to also predict the severity level of patients.

Keywords: COVID-19, diagnosis, prognosis, machine learning, metabolites, feature selection.

Contenido general

Índice de figuras

Índice de Tablas

Capítulo 1. Introducción

El COVID-19 es una enfermedad que surgió a finales del año 2019 debido a un nuevo tipo de coronavirus, denominado SARS-CoV-2, y que se expandió globalmente en pocos meses, causando una pandemia que alteró y afectó la vida de innumerables personas. A diferencia de otros coronavirus que afectan a los seres humanos, el SARS-CoV-2 continúa causando nuevos casos de enfermedad en todo el mundo hasta el día de hoy. Por ello, algunos especialistas opinan que es una nueva enfermedad con la que tendremos que convivir.

Durante las últimas décadas, se ha investigado el uso de la inteligencia artificial (IA) en una variedad de tareas, incluyendo el sector salud. Esto abarca el diagnóstico de enfermedades, el desarrollo y planificación de tratamientos, así como la atención médica de precisión, entre otros. Para el COVID-19, se ha investigado su uso en diversas tareas, destacando su aplicación en la detección mediante imágenes médicas de tórax, donde ha sido implementada y utilizada con éxito en varios centros de salud como una herramienta alternativa a las pruebas médicas tradicionales para diagnosticar la enfermedad.

Por otro lado, investigaciones han identificado alteraciones en diversas rutas metabólicas celulares causados por este virus. Se ha propuesto el uso de metabolitos como biomarcadores para el diagnóstico y pronóstico de la enfermedad. Sin embargo, debido a la gran cantidad de datos generados por las técnicas de la metabolómica, es necesario el uso de herramientas para su análisis. Recientemente, se ha propuesto el uso de la IA, en especial de su rama de aprendizaje automático (ML, por sus siglas en inglés), para el análisis de los datos generados por la metabolómica.

Por esta razón, en este trabajo se investiga el uso del aprendizaje automático, en combinación con los datos generados mediante métodos de la metabolómica, para su utilización como un método de diagnóstico y pronóstico de la enfermedad de COVID-19.

1.1 Antecedentes

COVID-19 es una enfermedad reciente causada por un nuevo tipo de coronavirus, denominado Síndrome Respiratorio Agudo Severo Coronavirus 2 (SARS CoV 2), la cual se

caracteriza por síntomas como tos seca, dolor de garganta y músculos, fiebre alta, fatiga y problemas respiratorios que aparecen dentro de los siguientes 2 a 14 días después de ser infectado (Ali & Alharbi, 2020; Shahin et al., 2022). A mayo de 2024, ya existe un total de contagios confirmados de más de 775 millones, y ha causado la muerte de más de 7 millones de ellos, de acuerdo con los datos publicados por la OMS, debido en gran parte a su alta tasa de transmisión, la cual tiene lugar mediante vías respiratorias al tener contacto con personas infectadas (Jamil et al., 2020).

A pesar de los grandes esfuerzos realizados por diversos sectores como el de la salud para tratar a los enfermos, o el político para el manejo de la pandemia, así como el uso de mascarillas, aplicación de aislamientos masivos y restricciones de movilidad, esta pandemia ha desafiado a diferentes sectores en todo el mundo: la llegada de esta nueva enfermedad puso a prueba no solo al sector salud, sino que alteró otros ámbitos como la economía global, el transporte, al sistema de comunicación y educación (Shahin et al., 2022). Dado el estado actual de la enfermedad, así como de la evolución que esta ha tenido durante estos años, los expertos en epidemiología opinan que el SARS-CoV 2 se continuará expandiendo globalmente durante varios años más (Baiges-Gaya et al., 2023).

Desde su aparición, los científicos e investigadores han trabajado de diferentes formas para el tratamiento tanto de la enfermedad como de la pandemia, utilizando diferentes enfoques y herramientas. Entre ellas, destaca el área de la Inteligencia Artificial (IA), y en especial sus ramas de aprendizaje automático, más conocida como *Machine Learning* (ML), y la de aprendizaje profundo o *Deep Learning* (DL) que han mostrado un alto potencial en el sector salud. Diferentes modelos de ML ya han sido estudiados y desarrollados por investigadores y científicos de datos para aplicaciones como el diagnóstico de COVID-19 (Hasan et al., 2021; T. Liu et al., 2022), predicción de complicaciones y requerimientos de hospitalización en una Unidad de Cuidados Intensivos (UCI) (Hasan et al., 2021; Moulaei et al., 2022; Noy et al., 2022; Saadatmand et al., 2022), y predicción de nuevos casos (Ghafouri-Fard et al., 2021), entre otros.

Dado el alto potencial mostrado por las áreas de ML y DL dentro del área de la salud para encontrar relaciones ocultas en los datos y realizar análisis de datos, y la capacidad para desarrollar sistemas para el diagnóstico, tratamiento, etc. (Davenport & Kalakota, 2019), surge el trabajo actual con el propósito de continuar con la investigación de la aplicación de estas áreas para la enfermedad de COVID-19. En este caso utilizando datos del perfil metabólico de

pacientes, para estudiar y proponer un modelo de ML que permitan diagnosticar y pronosticar la enfermedad en los pacientes de manera precisa.

El presente trabajo busca contribuir con herramientas y modelos alternativos a los existentes para el diagnóstico de esta enfermedad y su pronóstico, además de servir como base para posibles investigaciones futuras sobre la forma en la que el virus afecta al cuerpo humano, y en especial a la forma en la que este altera los procesos metabólicos que se llevan a cabo en las células de los seres humanos: debido a que las infecciones virales requieren la maquinaria metabólica del huésped para la síntesis de sus propios ácidos nucleicos, proteínas, lípidos y carbohidratos, y para obtener energía para la replicación viral (Camps et al., 2021), estas provocan importantes alteraciones en las rutas metabólica en el organismo infectado que son necesarias de conocer.

1.2 Planteamiento del problema de investigación

A febrero de 2023, el promedio de casos nuevos diarios de COVID-19 (tomado de los últimos 7 días) rebasa los 29 mil, de acuerdo con los datos reportados por la OMS. A pesar de los esfuerzos realizados hasta hoy, esta nueva enfermedad sigue desafiando a la humanidad de diferentes maneras, y sigue causando daños tanto humanos como materiales.

Aunque ya se han estudiado y propuesto algunos tratamientos contra esta enfermedad como el molnupiravir, la fluvoxamina y el paxlovid, que de acuerdo con varios estudios (Pourkarim et al., 2022; Wen et al., 2022) reducen la tasa de mortalidad y hospitalización en aproximadamente un 69%, estos medicamentos aún requieren de una mayor investigación y prueba. Así mismo, diferentes vacunas ya se han aplicado globalmente desde finales de 2021, las cuales han mostrado una eficiencia alta contra el COVID-19, mayor al 90% a los dos meses de su primera dosis, y de más de 70% a los 7 meses (Lin et al., 2022). Sin embargo, se han encontrado diferentes desafíos con respecto a la vacunación, como problemas de distribución en algunas regiones del mundo, o la aparición de nuevas variantes del virus como la Delta detectada en India en Octubre del 2022 y que se esparció a más de 140 países (Yu et al., 2021), debido en parte a la reducción de la eficacia de las diferentes vacunas en contra de esta variante, entre otros.

Por lo tanto el diagnóstico temprano de la enfermedad sigue siendo fundamental para combatir la enfermedad y de evitar la transmisión del virus, y por lo tanto también la cantidad de mutaciones

de este, especialmente considerando que más de un 80% de casos positivos son asintomáticos o presentan únicamente síntomas leves, y menos de un 20% presentan síntomas de moderados a graves (Jamil et al., 2020), pero que pueden llevar a la muerte, lo que causa que una gran cantidad de enfermos pasen desapercibidos pero que causen el contagio de más personas y, por consiguiente, también aumenten la probabilidad de nuevas mutaciones del virus.

Sin embargo, la aplicación rápida y masiva de pruebas ha tenido varios obstáculos, como su disponibilidad limitada, su exactitud en diferentes condiciones, y la logística para su distribución, entre otras, lo cual llevo al virus a expandirse rápidamente en todo el mundo debido a la limitada disponibilidad de pruebas y la baja tasa de detección de casos positivos, especialmente durante los primeros meses de la pandemia. Por otro lado, el método más utilizado como diagnóstico es la reacción en cadena de la polimerasa de transcripción inversa (RT-PCR) el cual, de acuerdo con algunos autores (Shahin et al., 2022), tiene una eficacia baja para detectar los casos positivos, y por lo tanto una baja sensibilidad para proceder con el tratamiento de los pacientes en fases tempranas: de acuerdo con la efectividad reportada (Davis et al., 2022), estas pruebas (RT-PCR) tienen una eficacia de hasta un 93% si son realizadas en el líquido de lavado bronco alveolar, pero esta se ve reducida al utilizarse en espunto e hisopos nasofaríngeos al 72% y 63%, respectivamente, y hasta el 32% en hisopos faríngeos o 29% en heces. En cambio, para otras pruebas sus valores de exactitud, sensibilidad y especificidad son poco conocidas.

Actualmente, el principal enfoque para el uso de ML y DL como un método de diagnóstico de COVID-19 alternativo ha sido mediante el uso de imágenes médicas tomadas por Tomografía Computarizada (TC) y rayos X de tórax: debido a que COVID-19 afecta principalmente a los pulmones (T. Liu et al., 2022)**,** estos modelos han logrado tener buenos desempeños, con medidas de exactitud cercanas o incluso superiores al 95% (T. Liu et al., 2022; Mishra et al., 2021; F. Zhang, 2021). Así mismo, otros enfoques han utilizado modelos con diferentes características como síntomas (Ali & Alharbi, 2020), señales de audio (Hemdan et al., 2022), etc.

Un enfoque muy prometedor que ha surgido más recientemente para el tratamiento de las enfermedades, como herramientas de diagnóstico y pronóstico, ha sido el uso de la metabolómica, el cual permite además obtener información sobre las interacciones huésped-patógeno a niveles de pequeñas moléculas (Hasan et al., 2021). Por ejemplo, se han analizado los metabolitos que podrían utilizarse como biomarcadores para el diagnóstico temprano del cáncer de mama (Subramani et al., 2022) y para el diagnóstico de la enfermedad celíaca (Ryan et al., 2015). Por

ello, algunos autores (Bardanzellu et al., 2022) han propuesto que la combinación de la metabolómica, microbiómica y el aprendizaje automático (ML), a lo que denominan las 3 M's, podrían ser alternativas óptimas y precisas en contra de enfermedades, en aplicaciones como el diagnóstico, evaluación y estratificación de riesgos, así como el manejo y toma de decisiones sobre los pacientes, dando paso a la medicina de precisión.

Actualmente existen ya algunos estudios sobre la relación de los metabolitos con COVID-19. En (Páez-Franco et al., 2021) por ejemplo, identificaron entre 50 y 60 metabolitos alterados en pacientes con COVID-19, algunos de ellos asociados con la severidad de la enfermedad o la variante del virus, mientras que en (Hasan et al., 2021) encontraron también diferentes metabolitos relacionados con la presencia de COVID-19 y el fallecimiento de pacientes. Aunque la literatura sobre la variación de metabolitos en pacientes con COVID-19 es muy amplia, la investigación de la combinación de esta con el aprendizaje automático aún es muy escasa.

Por estas razones, se plantea el uso de la metabolómica, combinado con el ML, como una posible herramienta para el diagnóstico y el pronóstico de esta enfermedad, con el propósito de generar modelos de ML que, combinados con la correcta combinación de mediaciones de metabolitos en los pacientes, tengan un desempeño comparable con los métodos utilizados actualmente (especialmente con las pruebas RT-PCR y el uso de imágenes médicas de tórax). Con esto, el objetivo final es presentar un modelo de ML robusto y confiable, que pueda ser aplicado posteriormente como una herramienta médica para el COVID-19 en centros de salud, así como presentar evidencia de nuevas técnicas que permitan estar más preparados ante posibles futuros brotes de virus.

1.3 Justificación del problema de investigación

Como se mostró en las secciones 1.1 y 1.2, el COVID-19 sigue siendo un problema actual que requiere de más trabajo e investigación. A pesar de los avances logrados, éstos fueron insuficientes para tratar adecuadamente la enfermedad y la pandemia. Una de las principales razones de la propagación del virus en todo el mundo fue la baja tasa de detección de casos positivos en fases tempranas, debido en parte a la baja disponibilidad de pruebas.

Además, se reconoce que el COVID-19 se ha convertido en una enfermedad más que debemos

enfrentar y prevenir en la vida diaria: aunque este virus podría desaparecer con el tiempo como lo hizo el SARS-COV, los datos reportados por la OMS y los continuos aumentos en el número de casos diarios muestran que no hay signos de que esto pase en el futuro cercano (Tiwari et al., 2023; Organización Mundial de la Salud [OMS], 2023). Por lo tanto, es importante contar con una amplia gama de métodos y alternativas para evitar o reducir sus efectos.

Los resultados de este trabajo de tesis podrían llevar a la creación y uso de nuevas herramientas para la detección y el pronóstico del COVID-19, lo que permitiría aumentar la cantidad de centros de salud capaces de diagnosticar la enfermedad y hacer pronósticos precisos. Además, se busca presentar evidencia del potencial de los modelos de aprendizaje automático (ML) y la metabolómica como alternativas confiables y precisas para el cuidado de la salud, lo que podría impulsar la investigación en otras aplicaciones, incluso para futuras infecciones y/o pandemias, lo que nos permitiría estar más preparados y actuar más rápidamente: diversos autores (Zumla et al., 2016; Sainté et al., 2015) han discutido ya la naturaleza y las implicaciones de las amenazas infecciosas emergentes, y han abordado las políticas y estrategias necesarias para mitigar su impacto, incluyendo la investigación de nuevas metodologías y el desarrollo de métodos de diagnóstico, vacunas, terapias, etc.

Finalmente, este trabajo puede servir como base para futuras investigaciones del COVID-19 desde la perspectiva de la metabolómica, lo que permitiría comprender mejor cómo la enfermedad afecta al cuerpo humano y a los procesos metabólicos a nivel celular, dando paso a posibles investigaciones futuras con enfoques explicativos.

1.4 Preguntas de investigación

- ¿Qué algoritmos supervisados de ML podrían ser factibles para generar un modelo de ML para el diagnóstico y pronóstico del COVID-19 utilizando metabolitos como características de entrada?
- ¿Puede un modelo de ML, basado en mediciones de metabolitos como características, alcanzar una eficacia similar o superior en el diagnóstico de COVID-19 a las obtenidas por las pruebas RT-PCR y a los modelos de ML con imágenes médicas de Tórax?
- ¿Cuáles son los metabolitos y el modelo supervisado de ML que alcanza un mejor desempeño

(basado en métricas de exactitud, sensibilidad, especificidad y F1) para el diagnóstico y predicción de una enfermedad severa de COVID-19?

- ¿Existe evidencia de variaciones en los metabolitos seleccionados para el modelo de aprendizaje automático en este trabajo debido a la enfermedad de COVID-19?

1.5 Objetivo general

Implementar y evaluar modelos de aprendizaje automático para diagnóstico y pronóstico (riesgo de hospitalización por enfermedad grave, y muerte) del COVID-19 utilizando un conjunto de metabolitos óptimo como características de entrada.

1.6 Objetivos específicos

- Seleccionar un conjunto de algoritmos supervisados de ML para su uso en la detección y pronóstico del COVID-19 mediante metabolitos como características.
- Encontrar el conjunto de metabolitos que permitan un mejor desempeño (en términos de exactitud y F1) para el diagnóstico y pronóstico de COVID-19.
- Comparar el rendimiento de algoritmos supervisados de ML, mediante mediciones de metabolitos como características, para el diagnóstico y pronóstico de COVID-19 y determinar si es posible alcanzar rendimientos iguales o superiores a las principales técnicas de diagnóstico utilizadas actualmente.
- Comparar los resultados de los metabolitos encontrados para los modelos de ML en este trabajo con la información existente de alteraciones de metabolitos debido a la enfermedad y nivel de severidad del COVID-19.

1.7 Hipótesis

Es posible utilizar modelos de ML supervisados (como *Logistic Regression* (LR), *Support*

Vector Machines (SVM) y *Random Forests* (Rf)) para el diagnóstico, predicción de enfermedad grave y muerte, utilizando medición de metabolitos de los pacientes con desempeños comparables con los resultados existentes mediante TC y rayos X de tórax (en términos de exactitud, F1, sensibilidad y especificidad), y superiores a los valores reportados por Davis et al., (2020) y Vandenberg et al., (2021) con pruebas de PT-PCR.

El conjunto de metabolitos óptimos para el modelo de ML, coincide con los metabolitos encontrados por otros autores como Hasan et al., (2021) y Xiong et al., (2022) que están relacionados con la enfermedad de COVID-19 y la presencia de enfermedad severa y/o muerte por esta enfermedad.

1.8 Estructura de la tesis

Esta tesis está dividida de la siguiente manera: en el capítulo actual se presentaron los antecedentes de este trabajo de investigación y los resultados que se pretende encontrar. El capítulo 2 corresponde al marco teórico del trabajo, donde se expondrán las principales teorías y resultados previos de los que se parte. El capítulo 3 describe el modelo de investigación seguido, y se describe la propuesta de investigación. Finalmente, el capítulo 4 muestra los resultados obtenidos en los diferentes experimentos, y en el capítulo 5 se exponen las conclusiones y objetivos alcanzados, así como las limitaciones de este estudio.

Capítulo 2. Marco Teórico

En este capítulo se realiza una revisión de las teorías que sirven como base para el desarrollo del presente trabajo de investigación, así como de las herramientas necesarias para llevar a cabo el análisis de los datos. En la sección 2.1 se hace una descripción de las teorías base: en la parte 2.1.1 se describen brevemente lo que son los coronavirus; en la parte 2.1.2 se brinda una descripción de la metabolómica; y en la parte 2.1.3 se realiza una recapitulación breve del Aprendizaje Automático. En la sección 2.3 se realiza una revisión de los trabajos relacionados, y en la sección 2.4 se hace una comparación entre los trabajos relacionados y el trabajo realizado en esta tesis. Finalmente, en la sección 2.5 se describe el esquema o modelo de investigación, basado principalmente en los criterios descritos por Sampieri.

2.1 Descripción de teorías base

2.1.1 Coronavirus

Los coronavirus son una familia de virus ARN envueltos que albergan cadenas de ARN de sentido positivo y pertenecen a la familia *Coronaviridae* del orden de los *Nidovirales.* Estos virus se dividen en cuatro géneros distintos: los gama y delta coronavirus, que infectan principalmente a aves, y los alfa y beta coronavirus, que afectan principalmente a diversas especies de mamíferos, incluyendo a los seres humanos (Corman et al., 2018; Shereen et al., 2020).

Durante el siglo actual, los coronavirus han cobrado una gran importancia en la salud pública mundial, debido a su capacidad de generar enfermedades en los seres humanos y en diferentes especies de animales, que pueden ir desde leves hasta muy graves, e incluso provocar la muerte, los cuales han causado diferentes epidemias que han afectado a diferentes países en todo el mundo.

2.1.1.1 Características de los coronavirus

Los coronavirus son un tipo de virus de tamaño diminuto, entre los 65 a 125 nm, y se componen

de una envoltura lipídica que protege al material genético y a las proteínas virales, con picos muy prominentes que sobresalen de 16 a 21 nm de la envoltura del virus (Payne, 2017), y son los que le dan el nombre debido a su forma parecida a una corona (figura 1).

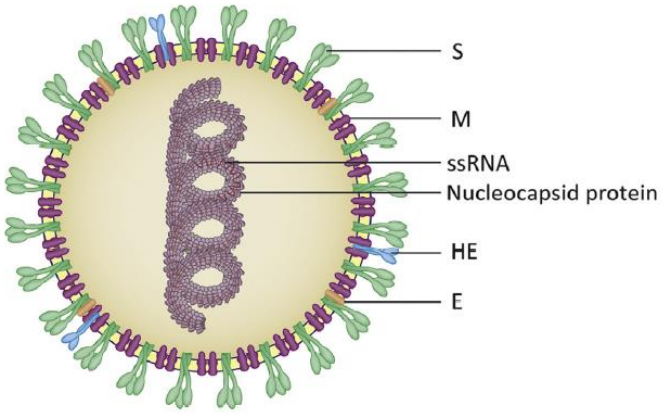

Figura 1. Estructura de los coronavirus capaces de infectar a los seres humanos. Tomada de (Shereen et al., 2020)

Los coronavirus son virus ARN de cadena positiva, que contienen un ARN monocatenario como material genético, de una longitud que oscila entre los 26 y los 32 kb, los cuales contienen entre 7 y 10 marcos abiertos de lectura (ORFs por sus siglas en inglés) (Payne, 2017), de los cuales el más largo codifica la glicoproteína de espiga (S), que es la responsable de la conexión y entrada del virus a las células huésped.

Todos los coronavirus codifican cuatro proteínas estructurales: S (espiga), M (membrana), E (envoltura) y N (nucleoproteína), de los cuales las tres primeras están asociadas a la membrana, y la cuarta es una proteína de la nucleocápside (Santos-López et al., 2021). Algunos beta-coronavirus aviares tienen una proteína de membrana con actividades hemaglutinantes y esterasas, denominada HE, que no se ha encontrado en coronavirus humanos (Payne, 2017). Las funciones de estas proteínas son:

- La proteína de Hemaglutinina-Esterasa (HE) es capaz de unirse a los receptores de ácido siálico, lo que ayuda al virus a unirse a ciertos tipos de células y tejidos.
- La proteína Espiga (S, por su nombre en inglés) es la más antigénica y externa, la cual le da la

forma característica de corona al virus, y es el medio mediante el cual ingresa en las células huésped. En el caso del SARS-CoV-2, esta proteína se una a la enzima convertidora de angiotensina 2 (denominada proteína ECA2) de la célula huésped.

- La proteína de Membrana (M) se encarga de dar estructura y estabilidad al virón.
- La proteína de Envoltura (E) participa en la formación de la envoltura del virus. Además, puede formar los canales iónicos para la liberación de los viriones de la célula huésped.
- La Nucleocápside (N) interviene en la síntesis del ARN viral, y se encarga de empaquetar y dar protección al genoma.

Además, los coronavirus codifican al menos 16 proteínas no estructurales, conocidas como nsp1 a nsp16, cada una con diferentes funciones (Santos-López et al., 2021).

2.1.1.2 Tipos de Coronavirus que infectan humanos (HCoVs)

Los coronavirus son conocidos desde la década de 1930, especialmente en aves y mamíferos como los murciélagos. Sin embargo, hasta 2002 estos virus se consideraban agentes patógenos de menor importancia para los seres humanos.

Los primeros descubrimientos de coronavirus capaces de infectar a los seres humanos (HCoVs) datan de la década de 1960, cuando se identificaron el HCoV-OC43 y el HCoV-229 (Hamre & Procknow, 1966; McIntosh et al., 1967), y posteriormente, en 2004 y 2005 fueron descubiertos el HCoV-NL63 y el HCoV-HKU1, respectivamente (Corman et al., 2018). A partir del 2002, se han identificado tres HCoVs con potencial epidémico:

- SARS-CoV, o síndrome respiratorio agudo severo coronavirus, fue identificado por primera vez en 2003 después del surgimiento de un brote durante el 2002 en la provincia de Cantón, República Popular China (Drosten et al., 2003). Este virus se expandió globalmente entre 2002 y 2004, afectando a más de 8 mil personas con una tasa de mortalidad de alrededor del 10% (Corman et al., 2018), y desde entonces no se han reportado nuevos casos.
- MERS-CoV, descubierto en 2012 debido a un caso de neumonía en Arabia Saudita, que afecto a más de 2 mil personas, con una tasa de mortalidad de aproximadamente el 35% (Corman et al., 2018). Desde entonces, se han presentado casos aislados o pequeños brotes de forma

esporádica, especialmente en la región del Medio Oriente.

- SARS-CoV 2, el más reciente, que surgió en Wuhan, providencia de Hubei, República Popular China, y desencadenó la pandemia de Covid-19 (enfermedad causada por este virus) (Shereen et al., 2020). Hasta mayo de 2023, ha habido más de 700 millones de casos confirmados y casi 7 millones de muertes a causa de esta enfermedad, de acuerdo a los datos reportados por la OMS.

2.1.1.3 Origen de los coronavirus en humanos

El origen de los coronavirus en los seres humanos aún no está completamente claro. Sin embargo, debido a la similitud genética encontrada con los coronavirus en animales, la línea más aceptada es que estos tienen un origen zoonótico (Cui et al., 2019; Singh & Yi, 2021), lo que significa que se originaron en animales, especialmente en murciélagos y roedores, y luego fueron transmitidos a los seres humanos a través de huéspedes intermediarios (figura 2).

De acuerdo a varios estudios filogenéticos, se ha propuesto que los CoV-NL63 y CoV-229E se originaron en murciélagos, mientras que los CoV-OC43 y CoV-HKU1 tienen su origen en roedores. Luego estos virus tuvieron un huésped intermediario con un amplio contacto con humanos, que en el caso del CoV-229E fue la alpaca y para el CoV-OC43 fue el ganado bovino, mientras que para los CoV-NL63 y CoV-HKU1 aún se desconoce (Corman et al., 2018).

Para el SARS-CoV, los investigadores inicialmente se enfocaron en los perros mapaches y las civetas de campo como el origen de estos virus. Sin embargo, investigaciones posteriores mostraron que estos animales solo eran huéspedes secundarios y que su origen podría haber sucedido en los murciélagos Rinolofo. En cambio, el MERS-CoV tiene como huésped intermediario a los camellos dromedarios, los cuales tienen un alto contacto con humanos en la región del Medio Oriente. Posteriormente, se sugirió también que estos virus son originados en los murciélagos, debido a que fue encontrado en las especies de murciélagos Pipistrellus y Perimyotis (Santos-López et al., 2021).

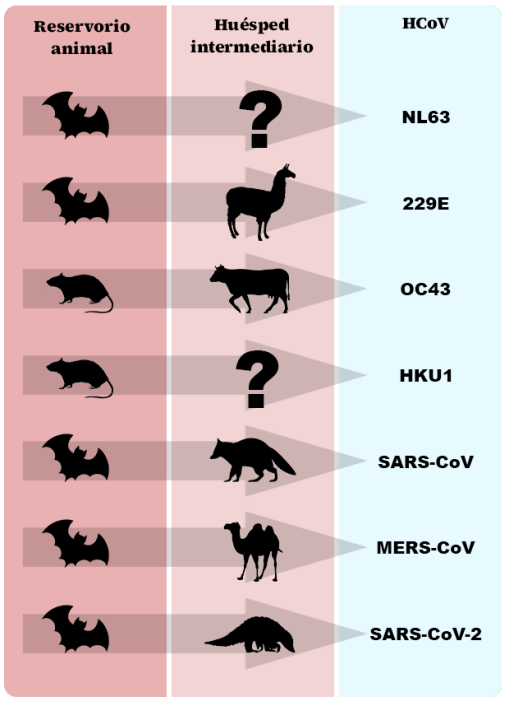

Figura 2. Diagrama del origen y huésped intermediario de los diferentes coronavirus encontrados en humanos. Adaptado de (Corman et al., 2018)

Los análisis genómicos revelaron que el SARS-CoV-2 está filogenéticamente relacionado con los SARS-CoV, por lo que desde el principio se propuse que los murciélagos podrían ser el reservorio primario (Santos-López et al., 2021; Singh & Yi, 2021). En particular, el genoma de este coronavirus tiene una similitud mayor al 96% a un coronavirus previamente hallado en murciélagos denominado Bat-CoV RaTG13. Estos dos coronavirus se diferencian principalmente en el dominio de unión al receptor S, y se ha propuesto que este virus y el Bat-Cov RatG13 pudieron haberse recombinado en algún punto debido a que esta secuencia es muy similar a la del coronavirus del pangolín malayo, por lo que el pangolín podría haber sido el huésped intermediario para el ingreso del virus a los seres humanos. Sin embargo, esta hipótesis no se ha aceptado completamente y sigue en discusión, debido a que se han encontrado otras secuencias en la proteína S que no existen ni en el Bat-CoV RaTG13 ni en el coronavirus del pangolín, por lo que huésped intermediario del SARS-CoV-2 aún es desconocido (Singh & Yi, 2021).

2.1.1.4 Transmisión de los coronavirus por contacto humano a humano

La transmisión de los coronavirus a través del contacto humano se realiza principalmente por la vía respiratoria (Ciotti et al., 2020). El proceso de entrada de los coronavirus en el cuerpo humano se puede describir en los siguientes pasos:

1. Contacto inicial: El virus ingresa al cuerpo humano debido al contacto cercano con una persona infectada, expuesta a partículas virales, como gotas respiratorias expulsadas mediante tos o estornudos o al tocar superficies contaminadas y luego tocarse la nariz, boca u ojos.
2. Unión de las células: la proteína de espiga (S) presente en la capa más externa del virus, donde se encuentra el Dominio de Unión al Receptor (RBD, por sus siglas en inglés), se une al receptor ACE2 (enzima convertidora de angiotensina 2) que se encuentra en las células huésped de los seres humanos.
3. Fusión y entrada: Una vez realizada la unión del virus al receptor ACE2, la membrana del virus se fusiona con la membrana de la célula huésped, liberando su material genético (ARN viral) en el citoplasma de la célula para iniciar la infección.
4. Traducción y replicación: El ARN viral liberado utiliza la maquinaria celular de la célula huésped para traducirse en proteínas virales, las cuales a su vez se replican para producir más proteínas virales.
5. Montaje y liberación: Las nuevas copias del virus se ensamblan en virones completos dentro de la célula, los cuales se liberan de la célula, a menudo dañándola o destruyéndola en el proceso. Posteriormente, estas partículas virales infectan a otras células propagando la infección.

2.1.1.5 Enfermedades por coronavirus en humanos

Los coronavirus humanos pueden causar una amplia gama de enfermedades, desde leves hasta graves, e incluso la muerte. Los síntomas más comunes de las infecciones por coronavirus incluyen fiebre, tos, dificultad para respirar, fatiga, dolor de cabeza y/o muscular y pérdida del olfato o el gusto. Otros síntomas menos comunes incluyen diarrea, vómitos, náuseas, conjuntivitis o exantema. Las complicaciones de las infecciones por coronavirus pueden incluir neumonía,

insuficiencia orgánica, problemas cardíacos y enfermedades neurológicas (Li et al., 2021). Además, existe evidencia de que algunos HCoVs como el SARS-CoV-2 pueden infectar el sistema nervioso central, causando enfermedades neurológicas como encefalitis, meningitis, encefalomielitis, polineuropatía, síncope o accidente cerebrovascular (Abdelaziz & Waffa, 2020). Sin embargo, aún es necesaria más investigación sobre estas complicaciones.

Los síntomas de las infecciones por HCoVs pueden variar en función del tipo de virus, la gravedad de la infección, la edad, la salud y las condiciones clínicas del paciente.

- Los síntomas más comunes de las infecciones por coronavirus endémicos clásicos, como el HCoV-229E, el HCoV-NL63, el HCoV-OC43 y el HCoV-HKU1, son fiebre, tos y rinorrea.
- Los síntomas de la COVID-19, causada por el SARS-CoV-2, son similares a los de las infecciones por coronavirus endémicos clásicos, pero también pueden incluir dificultad para respirar, pérdida del gusto o el olfato, dolor de cabeza, dolor muscular, escalofríos y fatiga.
- En casos graves, las infecciones por coronavirus pueden provocar neumonía, insuficiencia orgánica, problemas cardíacos y enfermedades neurológicas.

Estudios realizados por diversos autores han mostrado que los supervivientes del COVID-19 pueden presentar diversos síntomas incluso después de los 12 meses, a lo que se conoce como síndrome post-COVID-19 (Montani et al., 2022; Pierce et al., 2022). Los síntomas más comunes son fatiga, disnea, debilidad, dolor muscular, dolor de cabeza, trastornos del sueño, problemas cognitivos, problemas de la piel, problemas gastrointestinales y problemas cardiovasculares.

2.1.2 Metabolómica

La metabolómica es un área de las científica perteneciente a las ciencias ómicas encargada del estudio y cuantificación de los metabolitos presentes en sistemas biológicos. Su campo abarca una gran variedad de enfoques, como la cuantificación de estas pequeñas moléculas o el estudio de los procesos químicos relacionados a estos. La metabolómica busca comprender la complejidad de estas moléculas, sus interacciones que tienen en las rutas metabólicas en seres vivos, y cómo sus niveles son alterados debido a diferentes condiciones fisiológicas, patológicas o ambientales.

En esencia, la metabolómica genera un perfil de pequeñas moléculas (metabolitos) que se derivan

de un mecanismo metabólico celular y que refleja el resultado de una red compleja de reacciones químicas, proporcionando información sobre la fisiología celular (X. Liu & Locasale, 2017).

2.1.2.1 Metabolitos

La metabolómica implica la medición de metabolitos que se encuentran en un organismo y que constituyen su metaboloma, los cuales son moléculas pequeñas tanto endógenas (producidas de forma natural por el organismo que incluyen aminoácidos, ácidos orgánicos, ácidos nucleicos, azúcares, vitaminas, cofactores, entre otros) como exógenas (que no son producidos dentro del organismo, como medicamentos o aditivos alimentarios), generadas en sistemas biológicos durante las reacciones químicas asociadas al metabolismo de alimentos, fármacos, compuestos químicos, entre otros (Hernández Melo, 2021). Sin embargo, la metabolómica no está definida por ningún experimento en particular, sino que involucra un estudio del metabolismo de una manera integral (X. Liu & Locasale, 2017).

Los metabolitos son de vital importancia para los organismos vivos, ya que se encargan de controlar y regular una gran variedad de procesos biológicos, enviando señales químicas que comunican el estado fisiológico de una célula o un organismo. Sus niveles y patrones de expresión reflejan cambios tanto externos como internos de un sistema biológico, por lo que el estudio y análisis de estos metabolitos refleja directamente la actividad de las rutas metabólicas que conducen a su producción, proporcionando información esencial para comprender la fisiología, la salud y las respuestas adaptativas de los seres vivos que permitan comprender el estado biológico del sistema en cuestión (X. Liu & Locasale, 2017).

2.1.2.2 Enfoques de la metabolómica

La metabolómica utiliza tres enfoques para obtener el perfil metabólico de un paciente:

- La metabolómica dirigida se centra en un conjunto específico de metabolitos que han sido previamente seleccionados para su análisis, lo que requiere un conocimiento amplio previo y técnicas analíticas altamente sensibles y específicas para la detección y cuantificación del conjunto de metabolitos seleccionado, con lo que se obtienen las concentraciones absolutas de

las moléculas o las tasas de conversión de una molécula en otra.

- En la metabolómica no dirigida se monitorean miles de características desconocidas para identificar un amplio espectro de moléculas presentes utilizando técnicas no selectivas y de alto rendimiento, lo cual puede ser útil para identificar nuevos metabolitos presentes en un organismo o una ruta metabólica.
- Un experimento de la metabolómica puede ser semidirigido combinando elementos de los dos enfoques anteriores; en este caso, un conjunto de metabolitos es identificado y medido, sin embargo, es considerado semidirigido ya que no se cuenta con ninguna hipótesis, y por lo tanto no se excluye la posibilidad de descubrimientos no planificados.

2.1.2.3 Métodos de la metabolómica

La metabolómica hace uso de una gran variedad de métodos y herramientas que le permiten separar y cuantificar los niveles de metabolitos en una muestra biológica, así como de analizar e interpretar los resultados obtenidos. Aunque estas técnicas varían de acuerdo con los objetivos específicos del estudio, en general el proceso que sigue un estudio de la metabolómica puede ser dividido en las fases mostradas en la figura 3.

Figura 3. Metodología general para un estudio metabolómico.

1. **Recolección de las muestras.** Se obtienen muestras biológicas de las personas o los seres que se están estudiando. El tipo de muestras a recolectar es seleccionado debido al diseño experimental y al objetivo de estudio de la investigación, y pueden incluir sangre, orina, saliva o tejidos, entre otros.
2. **Preparación de las muestras.** Se realiza una preparación de las muestras obtenidas para

ser analizadas posteriormente, con el objetivo de cesar instantáneamente el metabolismo y minimizar la formación o degradación de metabolitos. Se realiza extracción con solventes orgánicos, entre otros para extraer de la muestra la mayor cantidad de metabolitos posibles para su determinación (Bourgin et al., 2023; Dettmer et al., 2007).

3. **Separación de los componentes de la mezcla.** Para analizar la composición de la muestra es necesario realizar primero una separación de los metabolitos, que se logra empleando típicamente técnicas de cromatografía como la Cromatografía Líquida (LC) o la Cromatografía de Gases (CG). La primera técnica se basa en la separación de la muestra utilizando una fase móvil líquida que se desplaza a través de una fase estacionaria, mientras que en la segunda la muestra es vaporizada e inyectada en una columna cromatográfica que separa a los componentes en función de su volatilidad y afinidad por la fase estacionaria.

4. **Determinación de la composición molecular.** Una vez separada la muestra se utilizan técnicas avanzadas que permiten determinar su composición molecular, comúnmente mediante Espectrometría de Masas (EM) o mediante Resonancia Magnética Nuclear (RMN). En el primer caso, la muestra se ioniza para generar moléculas cargadas, las cuales son detectadas y mostradas como un espectro donde el eje x contiene la relación masa-carga (m/z) y en el eje y la intensidad, mientras que con la segunda técnica (RMN) se emplea un campo magnético para medir la interacción de los núcleos atómicos de los metabolitos, y con ello determinar su estructura molecular. A partir de estas técnicas, se obtienen los datos de los niveles de metabolitos presentes en la muestra.

5. **Análisis de los resultados.** Los datos obtenidos de los niveles de metabolitos son analizados e interpretados utilizando herramientas como la estadística y el análisis de datos. Debido a la gran cantidad de información que es extraída de cada muestra, durante los últimos años se han venido empleando ampliamente diferentes técnicas de la inteligencia artificial, y en particular del aprendizaje automático, los cuales permiten encontrar estructuras o relaciones en el conjunto de entrada que con otros métodos sería muy complicado o imposibles de obtener.

2.1.3 Inteligencia artificial

La inteligencia artificial (IA) es una disciplina que se encuentra dentro de la ciencia de la computación, cuyo objetivo es diseñar sistemas informáticos y algoritmos que imiten las capacidades cognitivas e intelectuales de los seres humanos, como la capacidad de entender o comprender situaciones, de resolver tareas, o de aprender, etc., y abarca una gran variedad de subcampos de acuerdo al propósito, como el aprendizaje automático, la robótica, los sistemas expertos o la lógica difusa, entre otros.

2.1.3.1 Aprendizaje Automático

El Aprendizaje Automático, también conocido ampliamente como Machine Learning (ML), es un área de estudio perteneciente a la Inteligencia Artificial (AI), enfocada en generar sistemas informáticos que aprendan a resolver alguna tarea específica a partir de un conjunto de datos de entrenamiento, los cuales son ingresados a un algoritmo de aprendizaje que busca encontrar la forma de resolver la tarea de forma autónoma. Debido a su capacidad de identificar patrones en los datos que son difíciles de detectar o programar por los humanos, este es un campo de estudio que ha experimentado un gran crecimiento durante las últimas décadas, en gran parte debido al aumento en la capacidad de cómputo y de almacenamiento de los datos, lo que permite el desarrollo de algoritmos más sofisticados, con aplicaciones en una gran variedad de tareas, como clasificación, regresión, segmentación, procesamiento de texto y audio, entre otras.

El proceso mediante el cual el algoritmo de ML genera el modelo para realizar la tarea asignada se denomina *aprendizaje,* el cual se realiza utilizando un conjunto de datos denominado *conjunto de entrenamiento.* Una vez que el modelo de ML ha sido entrenado para realizar una tarea determinada, se espera que pueda generalizar y llevar a cabo la tarea para nuevos datos que no se utilizaron durante el entrenamiento.

2.1.3.2 Metodología del Aprendizaje Automático

Aunque la metodología seguida para encontrar un modelo de ML capaz de realizar la tarea deseada varía debido a varios factores, en general puede ser resumida en las siguientes 7 etapas, mostradas en la figura 4:

Figura 4. Metodología utilizada en el aprendizaje automático para generar un modelo (o sistema) capaz de realizar alguna tarea específica.

1. **Estudio del problema.** Es necesario conocer en gran medida la tarea que se desea realizar utilizando el aprendizaje automático, con el fin de seleccionar tanto los algoritmos y métodos a utilizar como el conjunto de datos y sus características con los que el modelo será utilizado.
2. **Recolección de datos.** De acuerdo con la tarea a resolver mediante al aprendizaje automático, se realiza una recolección de datos o ejemplos que utilizará el algoritmo de aprendizaje para generar el modelo mediante el proceso de aprendizaje.
3. **Preprocesamiento.** Se aplican una serie de operaciones y pruebas a los datos para adecuarlos a los modelos, entre las que se incluyen pruebas de normalidad, escalamiento de los datos, imputación o eliminación de datos faltantes, manejo de valores atípicos, entre otros.
4. **Extracción de características.** A partir de los datos de entrada, se realiza una extracción de características que mejor describan a cada uno de las instancias o ejemplos.
5. **Selección de características.** Se selecciona un conjunto con las características que mayor rendimiento obtienen (de acuerdo a alguna métrica de evaluación) para la tarea específica.
6. **Entrenamiento del modelo.** Se utiliza algún algoritmo de aprendizaje para generar un modelo a partir de los datos o ejemplos de entrenamiento.
7. **Validación.** Se utiliza alguna técnica para validar el modelo y obtener una estimación de su rendimiento para la tarea que fue entrenada utilizando datos nuevos (diferentes a los del conjunto de entrenamiento), comúnmente mediante validación cruzada y prueba con datos ciegos.

2.1.3.3 Modelo de Aprendizaje Automático

Un modelo generado por un algoritmo de aprendizaje automático, representado en la figura 5, es un sistema que recibe un conjunto de datos de entradas denominados *características* en forma de un vector $\vec{x} = (x_1, x_2, \dots, x_N)$, al cual se le denomina *vector de características.* Cada conjunto de datos de entrada $\vec{x}$ corresponde a una *instancia*, a partir del cual se desea predecir o inferir su etiqueta. A partir del vector de características de la instancia, el modelo realiza una inferencia o predicción, $\hat{\vec{y}} = h(\vec{x})$, tratando de predecir la etiqueta real de la instancia, $\vec{y}$.

Figura 5. Representación de un modelo de ML

El conjunto de datos de entrenamiento corresponde a un conjunto de m instancias con sus respectivos vectores de características $\vec{x}^{(i)}$, y posiblemente sus etiquetas reales $\vec{y}^{(i)}$. El conjunto de datos es representado en forma de la matriz X de $m \times N$, donde cada instancia corresponde a una fila, dada por:

$$X = \begin{bmatrix} \vec{x}^{(1)} \\ \vec{x}^{(2)} \\ \vdots \\ \vec{x}^{(m)} \end{bmatrix}. \quad (1)$$

Así mismo, se construye la matriz de las etiquetas Y al acomodar la etiqueta de la i-ésima instancia en la i-ésima fila,

$$Y = \begin{bmatrix} \vec{y}^{(1)} \\ \vec{y}^{(2)} \\ \vdots \\ \vec{y}^{(m)} \end{bmatrix}. \quad (2)$$

2.1.3.4 Tipos de Aprendizaje

De acuerdo a la forma en la que se lleva el proceso de aprendizaje, los algoritmos de ML pueden ser clasificados como:

- Aprendizaje supervisado: En este tipo de algoritmos, el algoritmo recibe un conjunto de entrenamiento que contiene un conjunto de datos etiquetados (el conjunto contiene tanto con los datos de entrada del modelo como con las salidas deseadas). El algoritmo genera un modelo capaz de predecir las salidas deseadas cuando se le introducen los datos de entrada para el conjunto de entrenamiento.
- Aprendizaje no supervisado: Este tipo de algoritmos de ML recibe un conjunto de datos de entrenamiento no etiquetado (el conjunto contiene las entradas del modelo, pero no cuenta con las salidas deseadas). En este caso, el algoritmo genera algún modelo a partir de los patrones encontrados en los datos.
- Aprendizaje semi-supervisado: Este tipo de algoritmos se utiliza cuando el conjunto de entrenamiento contiene solo una parte de los datos etiquetados, generalmente debido al alto costo de etiquetarlos.
- Aprendizaje por refuerzo: Este tipo de algoritmo de ML se utiliza para generar modelos capaces de tomar decisiones a partir de su entorno. En este caso, se le presenta al algoritmo un problema y se le permite experimentar con diferentes soluciones. A medida que el algoritmo interactúa, recibe recompensas o castigos de acuerdo a su desempeño, y este ajusta sus decisiones para maximizar la recompensa.

Para este trabajo, debido a que el objetivo es encontrar un modelo de aprendizaje automático para

diagnóstico y pronóstico de covid-19, se utilizará un algoritmo de ML para resolver un problema de clasificación (positivo vs negativo, enfermedad leve vs enfermedad grave). Así mismo, los datos cuentan tanto con las características (niveles de metabolitos) como con las etiquetas deseadas (positivo o negativo, y severidad de la enfermedad), por lo que se investiga el uso de algoritmos de aprendizaje supervisados.

2.1.3.5 Aprendizaje automático para problemas de clasificación

Los algoritmos de ML han probado ser una de las tecnologías más prometedoras para problemas de clasificación, y comúnmente es realizado mediante aprendizaje supervisado; es decir, el algoritmo de aprendizaje con el que se entrena al modelo recibe un conjunto de datos que contiene las características de las instancias de entrenamiento, y la clase real a la que pertenecen estas, y que se espera que sean predichas por el modelo.

En un problema de clasificación, el modelo de ML recibe un vector de características $\vec{\boldsymbol{x}}$ que representa a la instancia a clasificar, y a partir de ellas le asigna aluna etiqueta y, que pertenece a un conjunto finito de categorías posibles. En particular, para un problema de clasificación binaria, donde solo existen dos clases posibles, las clases pueden ser representadas como 0 o 1 (o negativa y positiva, respectivamente), de manera que la etiqueta $\boldsymbol{y}$ únicamente puede pertenecer al conjunto $\{0, 1\}$.

La salida del modelo puede ser tanto la clase a la que pertenece la instancia, como la probabilidad de que pertenezca a cierta clase (comúnmente la clase positiva). Para este último caso, la salida del modelo será un valor entre 0 y 1, y entonces la clase se le asignará de acuerdo a algún umbral r_i (también entre 0 y 1), de manera que si la salida es mayor a este umbral se clasificará a la instancia como positiva, y de lo contrario se clasificará como una instancia negativa.

En un modelo perfecto, existirá algún umbral r_i de tal manera que la salida del modelo de todas las clases positivas será mayor a este valor, y la de todas las clases negativas será menor a este valor. Si obtenemos la distribución de probabilidad para las clases negativas y positivas separadas (figura 6), un modelo es mejor a medida que el solapamiento entre ambas distribuciones sea menor.

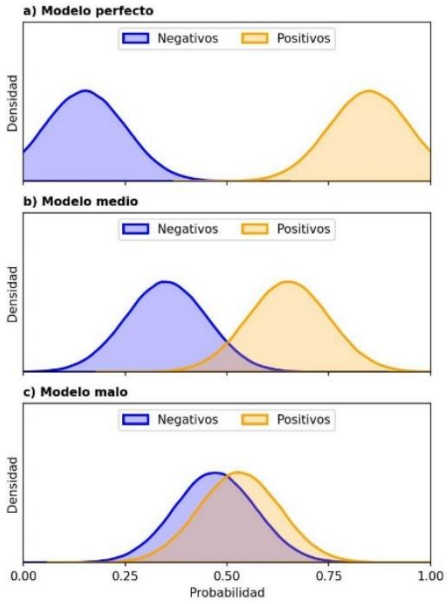

Figura 6. Distribución de probabilidad de las clases positivas y negativas para un modelo a) perfecto, b) medio y c) malo. Un modelo con mayor rendimiento es aquel con menor solapamiento entre ambas distribuciones.

2.1.3.6 Algoritmos de aprendizaje automático para la clasificación

El problema de clasificación es una de las tareas más comunes para las que es utilizado el aprendizaje automático, y por ello existen una gran variedad de algoritmos de clasificación. En particular, para el propósito de este trabajo se utilizarán los siguientes cuatro, debido a su bajo coste computacional, la interpretabilidad de estos modelos, y que en general son algoritmos que no tienden a sobre ajustarse a los datos:

- **Clasificador Bayesiano ingenuo**

  Los métodos de Bayes ingenuos son algoritmos de ML basados en el teorema de Bayes, los cuales asumen "ingenuamente" que cada para de características son independientes, dado el valor de la clase de la variable. Entonces, este algoritmo busca encontrar la probabilidad de que la instancia pertenece a la clase y^k, dado que tiene características $\vec{x} = (x_1, \dots, x_N)$, es decir $p(y^k|x_1, \dots, x_N)$. De acuerdo al teorema de Bayes, esta probabilidad está dada por

$$p(y^k|x_1, \dots, x_N) = \frac{p(y^k)p(x_1, \dots, x_N|y^k)}{p(x_1, \dots, x_N)} \quad (3)$$

Usando la asunción de independencia condicional de cada par de variables,

$$p(y^k|x_1, \dots, x_N) = \frac{p(y^k)\Pi_{i=1}^{N} P(x_i|y^k)}{p(x_1, \dots, x_N)} \quad (4)$$

Luego, la probabilidad $p(x_1, \dots, x_N)$ es constante dado el conjunto de entrenamiento, y la probabilidad $p(y^k)$ es la frecuencia relativa de la clase y^k en dicho conjunto, mientras que se asume que las probabilidades $P(x_i|y^k)$ tienen cierta distribución. En particular, para un clasificador gaussiano de Bayes ingenuo, se asume que esta última probabilidad tiene una distribución gaussiana, con media μ_y y desviación estándar σ_y obtenidas del conjunto de entrenamiento.

Por lo tanto, para una clasificación binaria la salida del modelo corresponde a la probabilidad de obtener la clase positiva y^1 dado el vector de características de la instancia $\vec{x}$.

- **Máquinas de Soportes Vectoriales**

El algoritmo de Máquinas de Soportes Vectoriales (SVM o SVC para un problema de clasificación, por sus siglas en inglés) busca encontrar alguna superficie (usualmente lineal) en el hiperplano N-dimensional definido por los vectores de características $\vec{x} = (x_1, \dots, x_N)$ del conjunto de entrenamiento que mejor separe a cada una de las clases, y a la vez que más se separe de los puntos en el hiperplano (figura 7).

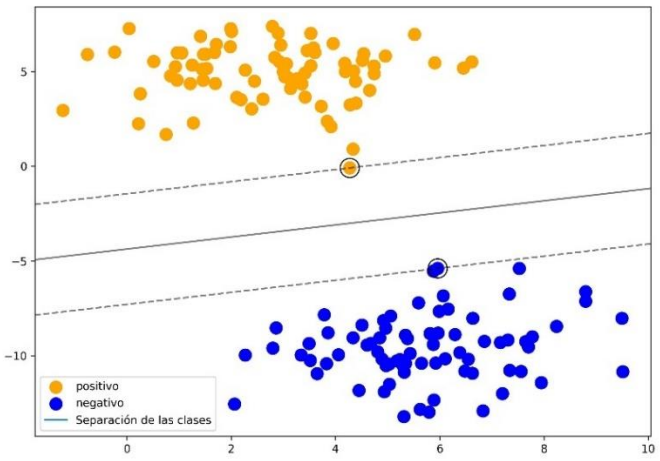

Figura 7. Separación de clases mediante una superficie (lineal) en el espacio de características, mediante máquina de soportes vectoriales.

Para el caso lineal (o kernel lineal), el hiperplano se puede escribir como

$$\vec{w} \cdot \vec{x} \; - \; b = 0, \tag{5}$$

y los parámetros $\vec{w}$ y b son obtenidos utilizando el conjunto de entrenamiento. Sin embargo, esta superficie que divide perfectamente a las clases solo es posible para conjuntos que sean linealmente separables, por lo que se puede especificar un hiperparámetro de regularización C para flexibilizar la construcción de esta superficie.

Adicionalmente, es posible utilizar kernels para separar clases que no sean linealmente separables.

- **Regresión Logística**

  La regresión logística (LR por sus siglas en inglés) es un algoritmo que permite resolver problemas de clasificación al encontrar la probabilidad de que la instancia pertenezca a la clase positiva. Para ello, supone que esta probabilidad depende del vector de características dado por

$$p(y^1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_N x_N)}} \quad (6)$$

Donde los coeficientes β_i son determinados realizando un ajuste mediante el conjunto de entrenamiento.

- **Bosques Aleatorios**

Los Bosques Aleatorios (RF por sus siglas en inglés) son una clase de algoritmos que están basadas en las técnicas denominadas ensambles, en las cuales se entrenan a varios modelos diferentes (en el caso de RF utilizando árboles de decisión) que en conjunto realizan la predicción mediante alguna técnica de agregación de las diferentes predicciones.

Un árbol de decisión comienza por un nodo raíz, y a partir de este surgen ramas que dirigen a otros nodos internos, conocidos como nodos de decisión, hasta llegar a los nodos terminales o nodos hoja (figura 8). Cada uno de estos nodos hojas pertenecen a alguna de las posibles clases del problema de clasificación. Inicialmente, el vector de características de la instancia es pasado al nodo raíz, y a partir de este, cada uno de los nodos de decisión dirigen a una de las ramas mediante alguna regla de decisión, hasta llegar a un nodo hoja. Finalmente, se asigna a la clase la etiqueta de la clase a la que pertenece esta hoja.

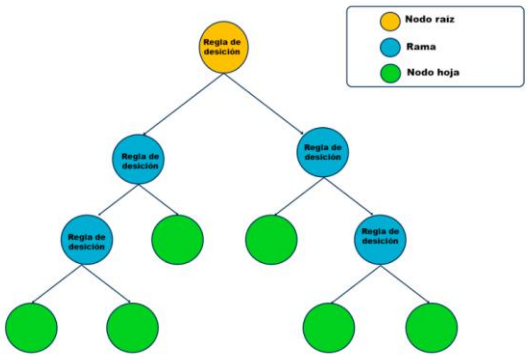

Figura 8. Representación de un árbol de decisión. Las características de la instancia a clasificar son pasadas en el nodo raíz, y a partir de aquí se va avanzando a los nodos subsecuentes mediante reglas de decisión hasta llegar a un nodo hoja, cada uno de los cuales tiene asignada una de las clases posibles.

En los bosques aleatorios se considera un conjunto de k árboles de decisión, y cada uno de ellos es entrenado con un subconjunto del conjunto de entrenamiento para generar un modelo de clasificación. La clasificación final realizada por el modelo puede ser tanto la clase más frecuente, posiblemente ponderada (clasificación dura), o el promedio de las probabilidades de cada clase (clasificación débil).

2.1.3.7 Selección de Características

Cuando los algoritmos de ML son aplicados a conjuntos de datos de altas dimensiones (alto número de características, comparado con el número de instancias), estos tienden a tener bajos rendimientos debido a un fenómeno causado por la lejanía entre las instancias en este espacio, conocido como la maldición de la dimensionalidad (en inglés, *curse of dimensionality*). Además, trabajar con estos conjuntos de datos suele ser poco apropiado debido a problemas como la dificultad para recolectar los datos, el alto costo tanto en memoria como en poder computacional para ejecutar los algoritmos, y la tendencia de estos a hacer sobreajuste.

El proceso de selección de características ha probado ser muy efectivo y eficiente para manejar conjuntos de altas dimensiones en el ML. El objetivo de la selección de características permite la creación de modelos más simples y comprensibles, además de mejorar su rendimiento.

La selección de características puede ser establecido como la búsqueda de un subconjunto de $\boldsymbol{d}$ características de un total de las $\boldsymbol{D}$ disponibles, que mejor desempeño logre al ser utilizado por un algoritmo de ML, de acuerdo con alguna función de interés. Sin embargo, a día de hoy no existe un método de FS capaz de encontrar este subconjunto óptimo en una complejidad computacional razonable, especialmente cuando el conjunto de datos contiene muchas características. En cambio, se han desarrollado diferentes métodos capaces de encontrar subconjuntos subóptimos.

- **Selección secuencial hacia adelante**

  El método de selección secuencial hacia adelante para el problema de FS consiste en agregar iterativamente una a una las características, e ir dejando en cada iteración esta nueva

característica siempre que se mejore la función criterio, o de lo contrario eliminarla. Inicialmente el subconjunto de características seleccionadas contiene 0 características, y se agrega en cada iteración una nueva característica: si el modelo mejora (o si se obtiene un resultado óptimo con la primera característica), esta se deja en el subconjunto seleccionado, de lo contrario esta se elimina.

- **Algoritmos genéticos**

Los algoritmos genéticos (o algoritmos evolutivos) son un grupo de algoritmos de optimizaciones estocásticos, los cuales siguen un conjunto de operaciones inspirados en el proceso natural de evolución, como apareamiento, herencia, etc.

En estos algoritmos cada posible solución es representado por un *individuo*, generalmente codificado como una cadena de bits (cada uno de ellos representando un *gen*), y un conjunto de individuos corresponde a una *población*. Inicialmente, se comienza con una población que consiste de un conjunto de individuos, comúnmente creados de forma aleatoria. Iterativamente (mediante *generaciones*), se genera una nueva población a partir de la anterior aplicando operadores genéticos de *cruzamiento* y *mutación* de forma estocástica, de forma que a medida de que pasan las generaciones (número de iteraciones), los individuos de la población tienden a tener un mejor desempeño que en las generaciones pasadas, obteniendo diferentes soluciones subóptimas del problema de optimización.

Para el problema de FS, si el conjunto de datos contiene $\boldsymbol{D}$ características, una posible solución (individuo) puede ser representado como una cadena de bits $\alpha_1, \alpha_2, \dots, \alpha_D$, donde cada uno de ellos representa si la característica pertenece o no (0 o 1, respectivamente) al conjunto. El operador genético de cruzamiento consiste en tomar dos individuos (padres), y generar un nuevo individuo, donde cada uno de los bits toma el valor de uno de los dos padres. La mutación se lleva a cabo en un pequeño grupo de individuos en cada generación, y consiste en intercambiar el valor de alguno de sus bits.

- **Boruta**

El método de selección de características de Boruta está basado en el algoritmo de Bosques Aleatorios, el cual fue implementado por primera vez para el lenguaje R, y utilizando una medida de la importancia de cada una de las características con las que se construye el modelo

como el valor Z, el cual se calcula encontrando la pérdida de exactitud del modelo al mezclar los atributos de las instancias para esta característica. El proceso seguido por este método, descrito por (Kursa, M. B., 2010), consiste de los siguientes pasos:

1. Se generan copias de cada una de las características y se agregan al conjunto de datos original (llamadas características sombra).
2. Para cada una de las características sombra, se mezclan los atributos de las instancias para eliminar cualquier correlación.
3. Se genera un modelo de clasificación utilizando el algoritmo de bosques aleatorios, y se calcula la importancia de cada una de las características utilizando los valores Z.
4. Se encuentra el puntaje Z más alto obtenido por las características sombra (denominado MZSA, por sus siglas en inglés),
5. Se utiliza una prueba estadística para comprar el puntaje Z de cada una de las características originales con el MZSA (el puntaje más alto alcanzado por las características sombra).
    - Si el puntaje Z de la característica original es significativamente menor al MZSA, entonces se elimina esta característica del conjunto de datos.
    - Si el puntaje Z de la característica original es significativamente mayor al MZSA, entonces se considera una característica importante, y se deja en el conjunto.
6. Se repiten los pasos anteriores hasta determinar todas las características importantes, o hasta alcanzar un límite preestablecido de iteraciones.

- **LASSO**

El método de Operador de Selección y Contracción Mínima Absoluta (LASSO, por sus siglas en inglés) es una técnica que funciona tanto para regularizar un modelo como para realizar selección de características. Este modelo toma como base a un modelo de regresión lineal, dada por

,

$$y_i' = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \quad (7)$$

donde $y_i{}'$ es la etiqueta predicha por el modelo para la i-ésima instancia, x_{ik}es la k-ésima característica de la i-ésima instancia, y los coeficientes β_k son los coeficientes de cada característica que son encontrados durante el entrenamiento del modelo, de manera que se minimice el error cuadrático medio de las predicciones. Es decir,

,

$$\hat{\beta} = \underset{\beta}{argmin}\left(\frac{(y_i - y_i')^2}{n}\right) \tag{8}$$

donde βes el vector de los coeficientes β_i, y $\hat{\beta}$ es el vector que minimiza el error de predicción. El método de LASSO consiste en agregar un término de penalización L1 a los coeficientes β, de manera que

,

$$\hat{\beta}(\lambda) = \underset{\beta}{argmin}\left(\frac{(y_i - y_i')^2}{n}\right) + \lambda||\beta|| \tag{9}$$

donde λ es un hiperparámetro del algoritmo de entrenamiento, que puede ser determinado mediante validación cruzada. Debido a propiedades geométricas, este término de penalización obliga a algunos de los coeficientes del modelo a hacerse cero, por lo que las características con coeficientes iguales a cero no son consideradas por el modelo, haciendo de forma indirecta una selección de características (Fonti, V., 2017). Entre mayor sea el valor de λ, se obtiene una menor cantidad de coeficientes no nulos.

Para un problema de clasificación, también es posible utilizar el modelo de regresión logística en lugar de la regresión lineal, por lo que en lugar de la ecuación 7 se utiliza la ecuación 6. Para este trabajo se utilizará esta última versión.

2.1.3.8 Evaluación

La evaluación de un modelo de ML consiste en estimar el rendimiento del modelo al desempeñar la tarea para la cual fue entrenado, comúnmente comparando las salidas obtenidas por el modelo, $\vec{\hat{y}} = \boldsymbol{h}(\vec{x})$, con las salidas reales o deseadas $\vec{y}$, y para ello existen una gran variedad de métricas.

Para una clasificación binaria, donde únicamente existen dos clases (comúnmente denominadas como negativa y positiva, o 0 y 1, respectivamente), se pueden definir el número de verdaderos positivos $\boldsymbol{T_P}$ (casos positivos que son clasificados correctamente), el número de verdaderos negativos $\boldsymbol{T_N}$ (casos negativos que son clasificados correctamente), el número de falsos positivos $\boldsymbol{F_P}$ (casos negativos que son clasificados incorrectamente) y el número de falsos negativos $\boldsymbol{F_N}$ (casos positivos que son clasificados incorrectamente), y con ellos se puede formar la matriz de confusión al acomodar los valores en una matriz de 2×2 como se muestra en la figura 9. De esta forma, un modelo perfecto tendrá valores no nulos en toda la diagonal principal de la matriz, y únicamente valores nulos en la anti-diagonal.

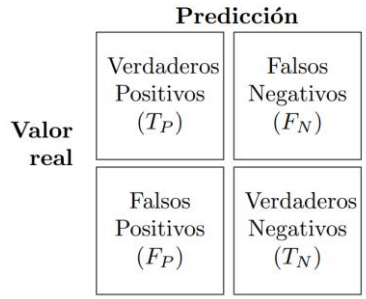

Figura 9. Matriz de confusión de la evaluación del modelo. Un modelo perfecto tendrá valores nulos en la anti diagonal de la matriz, y valores no nulos en la diagonal principal.

Además, a partir de los cuatro valores anteriores se pueden obtener las siguientes métricas:

$$\boldsymbol{Exactitud} = \frac{\boldsymbol{T_P} + \boldsymbol{T_N}}{\boldsymbol{T_P} + \boldsymbol{T_N} + \boldsymbol{F_P} + \boldsymbol{F_N}}, \tag{10}$$

$$\boldsymbol{Sensibilidad} = \frac{\boldsymbol{T_P}}{\boldsymbol{T_P} + \boldsymbol{F_N}}, \tag{11}$$

$$\boldsymbol{Especificidad} = \frac{\boldsymbol{T_N}}{\boldsymbol{T_N} + \boldsymbol{F_P}}, \tag{12}$$

$$\boldsymbol{Precisión} = \frac{\boldsymbol{T_P}}{\boldsymbol{T_P} + \boldsymbol{F_P}}, \tag{13}$$

que se pueden interpretar como la probabilidad de que el modelo clasifique correctamente a la instancia, independientemente de su valor real (exactitud); la probabilidad de que el modelo clasifique correctamente a una instancia, dado que su valor real es positivo (sensibilidad); la probabilidad de que la clasificación se incorrecta, dado que el valor real es positivo (especificidad); y la probabilidad de que la clasificación sea correcta, dado que el modelo clasificó a la instancia como positiva (precisión).

Para el caso en el que el número de instancias positivas y negativas en el conjunto sobre el que se realiza la evaluación sea muy diferente, la exactitud no es una métrica muy representativa del rendimiento del modelo. En tal caso, una alternativa es utilizar la exactitud balanceada, la cual se obtiene mediante la media aritmética de la sensibilidad y la especificidad, la cual es igual a la exactitud normal cuando el número de instancias es igual para cada una de las clases. O bien, se puede utilizar la métrica $\boldsymbol{F1}$, obtenida mediante la media armónica de la sensibilidad y la presición, y representa simétricamente tanto la precisión como la sensibilidad en una única métrica,

$$f1 = \frac{2}{\frac{1}{Precisión} + \frac{1}{Sensibilidad}}. \tag{14}$$

Debido a que todas estas métricas pueden ser interpretadas como una probabilidad, sus valores se encuentran entre 0 y 1, siendo 1 un modelo perfecto.

Si el modelo es capaz de hacer una clasificación débil (prediciendo las probabilidades de cada

clase, en lugar de una clase en particular), se puede variar el umbral de la probabilidad que divide a las predicciones de la clase positiva y negativa, obteniendo diferentes valores de sensibilidad y especificidad. Al graficar la $sensibilidad$ encontra del valor de $1 - especificidad$ al variar este umbral de probabilidad entre 0 y 1, se obtiene la curva ROC. El área bajo la curva ROC, denominado ROC AUC (por sus siglas en inglés), representa el grado o la medida en la que el modelo es capaz de distinguir entre las dos clases, siendo un modelo perfecto cuando es igual a 1, o un modelo aleatorio cuando es igual a 0.5; un valor menor a 0.5 indica que el modelo tiende a clasificar en la clase opuesta (clasifica a los 0's como 1's y viceversa).

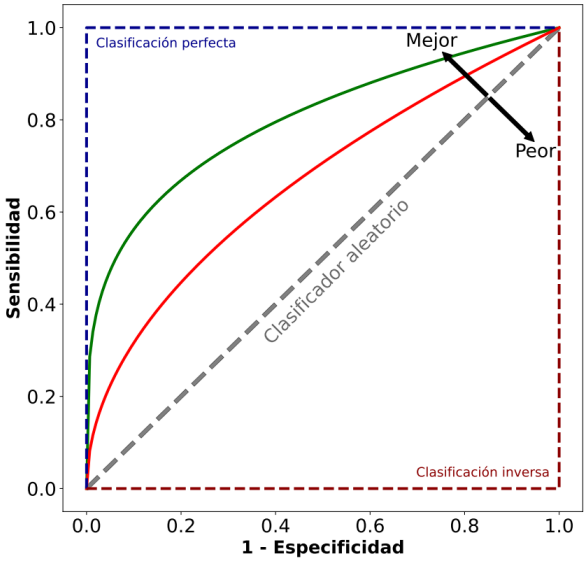

Figura 10. Curva ROC formada al variar el umbral de la probabilidad entre 0 y 1 a partir del cual una instancia es considerada como positiva. Un modelo perfecto tiene un AUC de 1, mientras que un modelo sin ninguna capacidad de clasificación obtendrá un AUC de 0.5. Una curva por debajo del clasificador aleatorio (AUC menor a 0.5) tenderá a clasificar a las instancias positivas como negativas y viceversa.

2.1.3.9 Validación

Durante el proceso de entrenamiento de un modelo de ML o DL, este se adapta a los datos utilizados para el entrenamiento, lo que se traduce en un alto rendimiento al evaluarlo con los mismos datos. Sin embargo, su desempeño comúnmente disminuye al enfrentarlo a nuevos datos, no utilizados durante el entrenamiento. Por ello, es de alta importancia utilizar un conjunto de

datos completamente diferente al de entrenamiento para su evaluación, conocido como *conjunto de validación*, que permite obtener una mejor estimación del rendimiento del modelo en un ámbito real.

La validación de un modelo consiste en utilizar el modelo previamente entrenado para la tarea que se desea realizar, para realizar una evaluación de este utilizando un nuevo conjunto de datos. Aunque se pueden obtener nuevos datos para realizar la validación, una vez que ya se ha generado el modelo, en la práctica es más común dividir todo el conjunto de datos con el que se dispone en los conjuntos de entrenamiento y prueba antes de comenzar con el procesamiento de los datos, procurando que los dos conjuntos tengan una estructura similar. Así, una parte de los datos es utilizada como conjunto de entrañamiento, mientras que los restantes se utilizan como conjunto de validación. La figura 11 representa este proceso de validación del modelo.

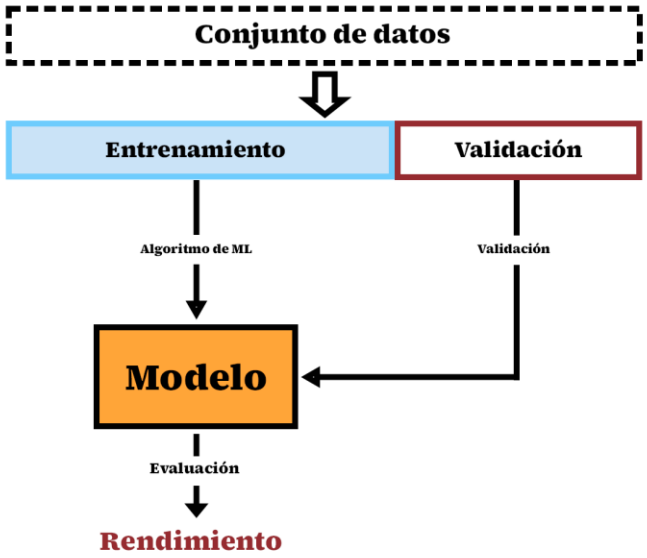

Figura 11. Para la evaluación del modelo, el conjunto original de datos se divide en los conjuntos de entrenamiento y de. El primero de ellos es utilizado para generar el modelo mediante algún algoritmo de ML, mientras que el segundo es utilizado para la validación.

Al realizar este procedimiento se obtienen dos valores del rendimiento del modelo: el rendimiento en el conjunto de entrenamiento y el rendimiento en el conjunto de validación. Si el rendimiento es bajo tanto en el conjunto de entrenamiento como en la validación, entonces se dice que el modelo hace un subajuste; si el rendimiento en el conjunto de entrenamiento es alto, pero este

disminuye considerablemente durante la validación, entonces se dice que el modelo hace un sobreajuste. Un buen modelo tiene un rendimiento alto en el conjunto de entrenamiento, y este no disminuye considerablemente en la validación.

Aunque con este proceso de validación se utiliza un conjunto de entrenamiento y uno de validación distintos, permitiendo obtener una mejor estimación del rendimiento del modelo en un ambiente real, el dividir el conjunto de datos en dos subconjuntos provoca una reducción en el número de datos disponibles para entrenar el modelo, y no todos los datos pasan por el proceso de validación. Una opción para resolver esto es realizar una Validación Cruzada (CV por sus siglas en inglés), representado en la figura 12. En esta técnica, se divide el conjunto de datos en N subconjuntos y se realiza el proceso de validación N veces: en cada validación, se utiliza uno de los subconjuntos de datos para la validación, y los $N - 1$ restantes para el entrenamiento, y utilizando algún método para promediar los rendimientos en cada uno de los subconjuntos. Si N es igual a la cantidad de instancias en el entrenamiento, de forma que en cada evaluación únicamente una instancia es predicha por el modelo para la evaluación, el proceso es llamado Validación Cruzada Dejando Uno Afuera (LOOCV por sus siglas en inglés).

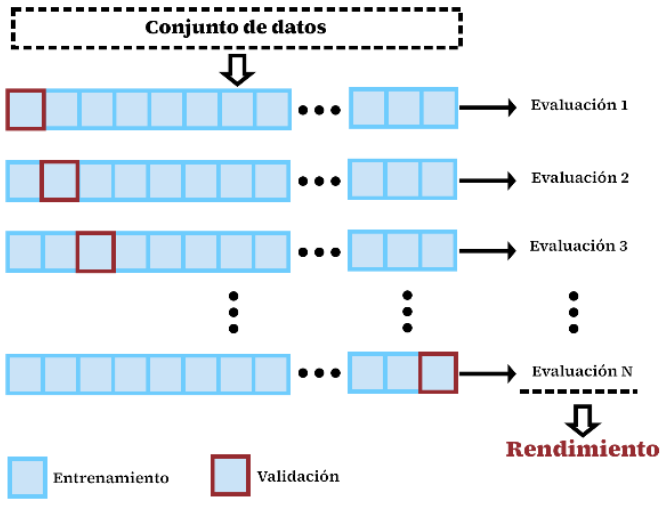

Figura 12. Validación Cruzada (CV). El conjunto de datos es divido en N subconjuntos de igual tamaño, y se realiza una validación para cada uno de ellos como conjunto de validación y los $N - 1$ restantes como conjunto de entrenamiento. El rendimiento del modelo se calcula como el promedio de las N evaluaciones.

2.1.3.10 Prueba a Ciegas

Para desarrollar un modelo de ML es común probar distintos algoritmos, cada uno con diferentes parámetros y configuraciones, para seleccionar el mejor modelo. Validar todos los modelos con el mismo conjunto de datos puede sesgar su rendimiento hacia ese conjunto particular y no reflejar su desempeño en situaciones reales.

La prueba a ciegas, representada en la figura 13, implica evaluar el modelo con datos diferentes a los usados para entrenar, validar y seleccionar el modelo. Por lo tanto, se divide el conjunto original en dos subconjuntos: uno para entrenar y evaluar los modelos y otro para probar el modelo seleccionado de forma ciega y obtener una mejor estimación de su rendimiento en situaciones reales. Este enfoque reduce el sesgo en el rendimiento del modelo y asegura una evaluación más precisa en escenarios prácticos.

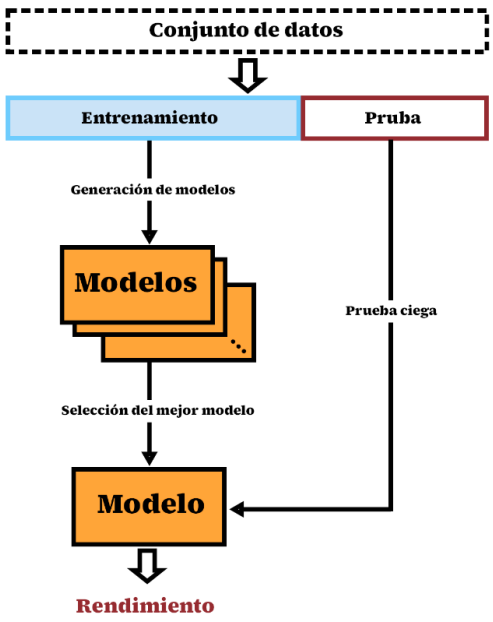

Figura 13. Proceso de selección y prueba ciega. A partir del conjunto de datos se obtienen dos subconjuntos, de entrenamiento y prueba. Una vez entrenados y evaluados los modelos con el primer conjunto mediante alguna técnica como validación cruzada, se selecciona el mejor de ellos, el cual es evaluado utilizando el segundo conjunto, el cual no ha sido

utilizado previamente.

2.2 Principales estudios relacionados

Actualmente varios autores ya han reportado sobre cambios de los niveles de metabolitos debido a la infección del COVID-19. Los estudios existentes realizados por diferentes autores (Danlos et al., 2021; Hasan et al., 2021; Páez-Franco et al., 2021; Shen et al., 2020; Xiong et al., 2022) han mostrado que esta enfermedad causa alteraciones en diversas rutas metabólicas. Diversos autores han realizado estudios sobre el cambio en los niveles de metabolitos debido a la presencia del COVID-19, y han sugerido su uso como biomarcadores para el diagnóstico y pronóstico de esta enfermedad, enfatizando además las ventajas de reducción del costo y el tiempo de respuesta de las pruebas en los laboratorios de microbiología de diagnóstico.

Específicamente, estos estudios mencionan principalmente la vía de la quinurenina, el metabolismo de aminoácidos, de lípidos y el metabolismo energético como las rutas metabólicas principalmente alteradas, y en particular, (Hasan et al., 2021) propone el uso del triptófano, la kynurenina, 3-hidroxiquinurenina, así como otros metabolitos de las rutas metabólicas mencionadas anteriormente como posibles biomarcadores para el diagnóstico de los pacientes, mientras que en (Valdés et al., 2022) proponen también a metabolitos relacionados con los ácidos grasos, los fosfolípidos, el triptófano, entre otros para diferenciar a pacientes con enfermedad en etapa terminal de pacientes sanos o en etapas tempranas .

En una investigación realizada al principio de la pandemia (Shen et al., 2020), se examinó el suero de 46 pacientes con COVID-19 y 53 individuos sanos. En este análisis, identificaron y cuantificaron 941 metabolitos, de los cuales 373 mostraron alteraciones significativas debido al COVID-19. Más de 200 de estos metabolitos estaban correlacionados con la gravedad de los pacientes y más de 100 lípidos que estaban regulados negativamente con el suero de los pacientes enfermos, probablemente debido a daños en el hígado, así como diferencias significativas entre los pacientes enfermos y los controles sanos en términos de metabolismo de purina, glutamina, leucotrieno D4 y glutatión.

Adicionalmente, en esta investigación llevada a cabo por (Shen et al., 2020) realizaron un modelo

de bosques aleatorios para la clasificación de pacientes severos de COVID-19 utilizando los niveles de 7 metabolitos y 22 proteínas, en el cual obtuvieron una exactitud del 93.5% en el conjunto de entrenamiento (18 pacientes no severos y 13 pacientes severos), y posteriormente una exactitud de 84.2% en la prueba (16 de 19 pacientes clasificados correctamente).

Otras investigaciones realizadas por diferentes autores han mostrado resultados similares. Por ejemplo, se han encontrado alteraciones en los metabolitos relacionados con los niveles de severidad de COVID-19 (Danlos et al., 2021), encontrando alteraciones del metabolismo del triptófano en la vía de la quinurenina con la inflamación e inmunidad en pacientes enfermos graves en comparación con enfermos leves, así como niveles elevados de quinurenina y niveles disminuidos de arginina, sarcosina y LPC en pacientes enfermos de COVID-19 en comparación con pacientes sanos. Así mismo, se ha caracterizaron las vías metabólicas de pacientes enfermos en comparación con controles sanos (Blasco et al., 2020), revelando alteraciones en las rutas metabólicas de la arginina, el triptófano o el metabolismo de las purinas.

2.3 Contribuciones y limitaciones de estudios previos

En la actualidad se ha logrado un considerable avance en la comprensión del SARS-CoV-2 y la enfermedad COVID-19 en su conjunto, gracias a la extensa investigación realizada desde la irrupción de este nuevo coronavirus hasta la actualidad. Específicamente, los estudios previos centrados en la metabolómica de pacientes afectados han desempeñado un papel crucial en la elucidación de los mecanismos moleculares subyacentes al COVID-19, avanzando cada vez más en el conocimiento sobre la forma en la que este virus altera los procesos metabólicos celulares de los pacientes enfermados y las rutas metabólicas involucradas en estos procesos, abriendo nuevas oportunidades para el desarrollo de nuevas terapias, así como la identificación de biomarcadores potenciales como una herramienta prometedora para el diagnóstico y la evaluación o pronóstico de la gravedad de la enfermedad. A pesar de que los estudios previos han incluido pacientes con diferentes características tanto clínicas como sociodemográficas, además de que utilizan distintas técnicas, los estudios concuerdan en que este virus desencadena alteraciones en diversas rutas metabólicas que pueden influir en la progresión y severidad de la enfermedad, entre las que se mencionan principalmente el metabolismo del triptófano, energético, de aminoácidos, de lípidos y

de nucleótidos.

Sin embargo, estas investigaciones aún se encuentran en curso y las conclusiones no están del todo claras. La heterogeneidad de la enfermedad y las condiciones de los pacientes estudiados, el tamaño reducido y sesgado de la muestra que se ha utilizado, las variables en las metodologías empleadas y la falta de estudios longitudinales para comprender los cambios en metabolitos y rutas metabólicas durante el curso de la enfermedad son algunas de las principales limitaciones que constituyen obstáculos que requieren abordarse con mayor profundidad. Por tanto, aún se requieren más estudios para comprender a fondo estas interacciones del virus en los pacientes.

Aunque se han obtenido resultados prometedores en el uso potencial de metabolitos como biomarcadores para desarrollar nuevas técnicas de diagnóstico y prever la gravedad de la enfermedad, así como en la aplicación de técnicas de aprendizaje automático para estas funciones, aún no existe consenso sobre qué metabolitos específicos o qué modelos técnicos ofrecen el mejor rendimiento. Por lo tanto, actualmente hay escasas o nulas implementaciones de estas técnicas.

Aunque se han realizado diversas investigaciones sobre la variación de metabolitos debido al COVID-19 y se ha demostrado su potencial como biomarcadores para el diagnóstico y pronóstico de la enfermedad, su enfoque principal ha sido descriptivo o explicativo. Actualmente, hay escasa investigación sobre la aplicación práctica de estos resultados. El objetivo principal de esta investigación es utilizar las alteraciones en los niveles de metabolitos en pacientes con COVID-19 para diagnosticar y clasificar la gravedad de la enfermedad mediante técnicas de aprendizaje automático. Por lo tanto, el trabajo se centra en la identificación de uno o varios metabolitos y en el desarrollo de un modelo que permita una aplicación práctica eficiente para estas tareas.

Por ello, a diferencia de los trabajos existentes cuyo enfoque principal se ha centrado en encontrar variaciones en niveles de metabolitos en pacientes enfermos de COVID-19 y en entender los procesos biológicos subyacentes, en este trabajo las principales etapas se centran en la búsqueda de un modelo práctico para desarrollar nuevas técnicas para el diagnóstico y pronóstico de la enfermedad mediante la aplicación de modelos de aprendizaje automático, con el fin de presentar evidencia del rendimiento de estas técnicas para su posible uso práctico como alternativas a las existentes.

2.4 Modelo o esquema general de investigación

De acuerdo con los criterios para catalogar la investigación (Sampieri, 2016), el diseño del presente trabajo es de carácter *no experimental*, ya que no se pretende realizar una manipulación intencional de las variables independientes, sino que se examinan los hechos tal como se han dado de forma natural, sin ser provocados intencionalmente: las variables independientes ocurren, y no es factible manipularlas por cuestiones éticas. Además, debido a que involucra el contagio por el SARS-CoV 2 en seres humanos y los posibles cambios en los niveles de metabolitos, no es posible desarrollar una investigación con enfoque experimental debido a cuestiones éticas. Así mismo, se trata de una investigación *transversal*, debido a que los datos fueron recolectados a partir de pacientes con COVID-19 y pacientes sanos en un periodo de tiempo dado, y no se busca encontrar los cambios con el paso del tiempo.

El alcance del trabajo de investigación es *correlacional* debido a que se pretende utilizar las técnicas del ML para encontrar relaciones o estructuras ocultas en los datos (C. Zhang et al., 2020) (niveles de metabolitos, resultados de test de COVID-19 y nivel de severidad de la enfermedad, y datos del paciente) que sirvan como técnicas de diagnóstico y pronóstico del COVID-19. Además, es *no explicativo*, debido a que no se busca atribuir causalidad.

Capítulo 3. Método y propuesta de investigación

En este capítulo se describe de forma detallada el método y la propuesta de investigación. La sección 3.1 describe de manera general el modelo de investigación, y en la sección 3.2 se describen de manera detallada cada una de las fases.

3.1 Modelo de Investigación

El diseño de investigación que se propone comprende diversas fases, las cuales se ilustran en la figura 14 y se describen más detalladamente en la sección 3.2, el cual es una adaptación de la metodología general descrita en la sección 2.1.3.2.

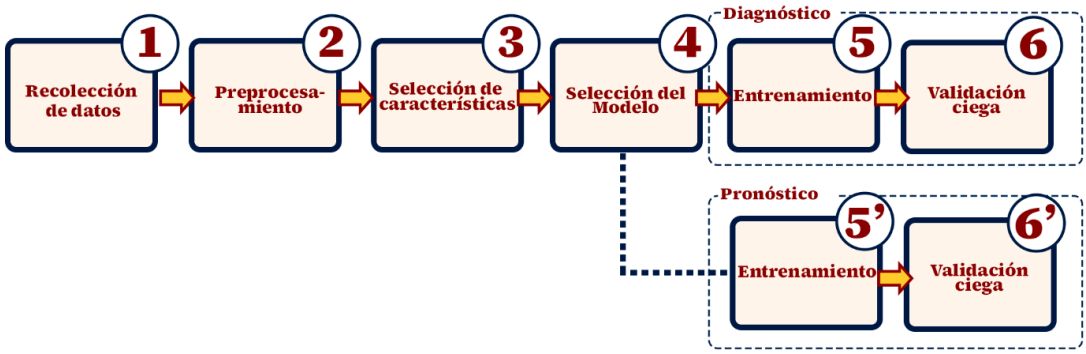

Figura 14. Etapas de la metodología de investigación propuesta.

La investigación comienza con una recolección de datos a partir de pacientes sospechosos de COVID-19, que posteriormente son clasificados en 4 diferentes grupos (el grupo G1 que consiste en los pacientes sanos, y los grupos G2, G3 y G4 con resultados positivos, clasificados según la escala de severidad propuesta por la OMS). La segunda etapa consiste en un preprocesamiento de los datos donde estos son preparados para su análisis utilizando las herramientas del aprendizaje automático (ML, por sus siglas en inglés). En la etapa 3 se realiza una selección de las características (metabolitos) que son más relevantes para generar un modelo de ML, mientras que en la etapa 4 se realiza la elección del mejor modelo de ML para llevar a cabo el diagnóstico de COVID-19. Esto incluye la selección del algoritmo óptimo y la búsqueda de sus hiperparámetros. Finalmente, el modelo propuesto es entrenado tanto para el diagnóstico (etapa 5) como para

pronóstico de enfermedad grave (etapa 5'), y es evaluado mediante una validación con datos ciegos (no utilizados previamente).

Debido a que se utilizan como características los niveles de metabolitos obtenidos directamente durante la recolección de datos, se omite la fase de extracción de características.

3.2 Descripción de la propuesta

A continuación, se describen detalladamente cada una de las fases de las que consiste la propuesta de investigación, representadas en la figura 14.

3.2.1 Recolección de los datos y selección de la muestra

El proceso de recolección de datos realizado por López, Y., et. al. (2020), consistió en la obtención de datos clínicos de 158 pacientes: 83 hombres y 71 mujeres (4 pacientes no especificados) de entre 30 y 70 años, posterior al ingreso al Instituto Mexicano del Seguro Social (IMSS) entre marzo y noviembre de 2020 (figura 15). Entre ellos, existen 37 personas sanas (grupo G1), y 121 pacientes fueron confirmados positivos: 41 fueron pacientes con una enfermedad leve (G2) y 80 pacientes fueron hospitalizados de acuerdo a la escala de severidad de la OMS [CITAR], de los cuales 35 fueron clasificados como moderados/graves (grupo G3), y 45 se incluyeron en los pacientes críticos (grupo G4).

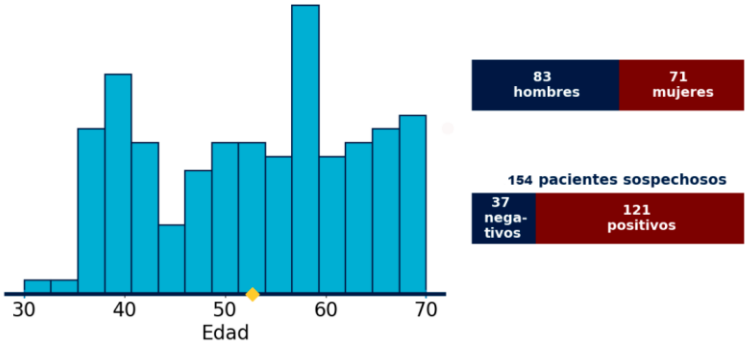

Figura 15. Distribución de los datos recolectaos.

Para la cuantización de los datos de laboratorio se recolectaron muestras de plasma dentro de los 2 días posteriores a la hospitalización y antes del diagnóstico de laboratorio mediante pruebas RT-PCR, antes de recibir tratamientos si fueron prescritos, de donde se cuantizaron 110 metabolitos y 13 citoquinas/quimosinas utilizando Cromatografía de Líquidos acoplada a la Espectrometría de masas (LC-MS/MS por sus siglas en inglés) y citometría de flujo, respectivamente, junto con otros datos sociodemográficos.

Todos los pacientes incluidos en la obtención de datos fueron informados por escrito sobre la recolección de sus muestras para fines de investigación, y se les otorgó el derecho a rechazar dichos usos, en conformidad con la declaración de Helsinki, aprobado por el Comité de Ética del IMSS.

Por lo tanto, de acuerdo con los criterios de (Sampieri, 2016). la selección de la muestra corresponde al tipo *no probabilístico,* ya que no se emplean técnicas de selección por métodos probabilísticos para la selección de los participantes o para la estimación del tamaño de la muestra. En su lugar, los datos corresponden a casos de interés de la investigación, y son seleccionados principalmente por la accesibilidad, por lo que no se busca generalizar los resultados de la muestra a la población. Sin embargo, en complemento con otros estudios similares existentes como los realizados por (Hassan et al., (2021), Xiong et al., (2022), entre otros, así como con estudios futuros, se busca aportar evidencia para llegar a generalizaciones.

El conjunto de datos recopilado originalmente consta de 156 variables o características. No obstante, para el propósito de este estudio se optará por utilizar exclusivamente las 110 mediciones de metabolitos disponibles. En este contexto, se estableció el criterio de inclusión-exclusión, limitándonos a trabajar únicamente con aquellas características que presenten menos del 30% de datos faltantes. Como resultado de esta selección, se eliminaron cuatro características, resultando en un conjunto de datos compuesto por 158 pacientes y 106 características (mediciones de metabolitos). Todas las características obtenidas por la prueba de laboratorio son de tipo *cuantitativas* con nivel de *razón*, mientras que el grupo al que pertenecen los pacientes es de tipo cualitativa con un nivel de medición *ordinal*

3.2.2 Preprocesamiento de los datos

La fase principal de este proceso se centra en la depuración y preparación de los datos para su posterior empleo por los algoritmos de Aprendizaje Automático. En este sentido, se genera un modelo de preprocesamiento de datos (pipeline) mediante el conjunto de entrenamiento, que consiste en el escalado de los datos mediante estandarización y la imputación de datos faltantes mediante un algoritmo de K Vecinos más Cercanos (KNN, por sus siglas en inglés). Es importante destacar que estos dos modelos son ajustados exclusivamente al conjunto de entrenamiento y, posteriormente, se aplican para realizar la imputación en el conjunto de prueba.

3.2.3 Selección de características

El propósito de esta etapa es la selección de un subconjunto de metabolitos (características) entre los 106 disponibles en el conjunto de datos. Esto tiene como objetivo la construcción de un modelo de aprendizaje automático más efectivo para el diagnóstico/pronóstico (predicción de enfermedad grave) de COVID-19. La idea central es identificar y utilizar exclusivamente aquellas características que sean más relevantes o informativas, eliminando las que podrían resultar redundantes o irrelevantes.

Esta estrategia tiene beneficios significativos. Por un lado, contribuye a la creación de modelos más eficientes y simples, mejorando así su rendimiento. También aborda los desafíos asociados con conjuntos de datos de alta dimensionalidad, comúnmente conocidos como la maldición de la dimensionalidad, causada por la dispersión de datos en espacios de alta dimensión. Además, esta fase proporciona información valiosa sobre las características más destacadas e informativas relacionadas con la enfermedad. Asimismo, facilita la interpretación de los modelos, lo que es crucial para comprender y aplicar eficazmente los resultados en el contexto clínico.

Debido a que actualmente no existe ninguna técnica que garantice encontrar el conjunto de características óptimo, se utilizarán 4 distintos métodos: selección secuencial hacia adelante, algoritmos genéticos, boruta y lasso, por lo que se propondrán y probarán 4 conjuntos de características distintos. Para el método de algoritmos genéticos se utilizará la librería de *sklearn-genetic-opt* versión 0.10.1, para el método de boruta el paquete *Boruta* versión 0.3 *(Kursa, 2010)*, y para el método de lasso el paquete *Scikit-Learn* versión 1.2;2 (Pedregosa, 2011), mientras que

para el método de selección secuencial hacia adelante se utilizará una implementación propia y que se puede ver en el repositorio del trabajo en LINK.

Para los modelos de algoritmos genéticos, boruta y lasso, se muestran los principales parámetros utilizados en las tablas del apéndice B, los cuales fueron seleccionados de manera arbitraria. Para los parámetros no especificados, se utilizaron los valores por defecto por cada uno de los paquetes utilizados, por lo que pueden ser revisados en la documentación de estos, en sus respectivas versiones.

3.2.4 Selección del modelo

Esta etapa comprende tanto la selección del algoritmo de aprendizaje automático para generar el modelo, como la búsqueda de sus hiperparámetros. Para ello, se seleccionan aquellos que sean interpretables, computacionalmente simples, propensos a tener un bajo sobreajuste y ampliamente utilizados en la literatura para tareas clínicas. Por lo tanto, se proponen los siguientes cuatro: clasificador gaussiano ingenuo, regresión logística, bosques aleatorios y máquinas de soporte vectorial (SVM, por sus siglas en inglés).

Cada uno de los cuatro algoritmos propuestos es probado con cada uno de los cuatro conjuntos de características propuestos en la fase anterior, por lo que en total se probarán 16 combinaciones posibles. Para cada uno de ellos, se realiza una búsqueda de hiperparámetros utilizando una búsqueda de cuadrícula con los valores mostrados en el apéndice C. Para la evaluación de cada posible combinación se utiliza una validación cruzada dejando uno afuera con el conjunto de datos de entrenamiento.

3.2.5 Entrenamiento

En esta etapa, se lleva a cabo la ejecución del modelo de aprendizaje automático propuesto y su correspondiente entrenamiento utilizando el conjunto completo de datos de entrenamiento con las características específicas seleccionadas.

Por un lado, se realiza el entrenamiento del modelo para llevar a cabo el diagnóstico de COVID-

19, clasificando entre los grupos G1 versus G2 + G3 + G4 (etapa 5). Por otro lado, el mismo modelo se implementa para predecir enfermedades graves, clasificando entre los grupos G2 versus G3 + G4 (etapa 5').

A pesar de que el modelo inicialmente se diseñó exclusivamente para el diagnóstico de COVID-19, se extiende su aplicación para realizar pronósticos. Esto se debe, por un lado, a la reducción del conjunto de datos al excluir al grupo G1, y por otro lado, a la decisión de utilizar un mismo tipo de modelo y conjunto de características para ambos propósitos.

3.2.6 Validación ciega

Durante esta etapa, el objetivo es evaluar el rendimiento final del modelo y su capacidad de generalización al emplearlo con datos nuevos. Este proceso se lleva a cabo utilizando datos no utilizados previamente, que han sido separados desde antes de la fase de preprocesamiento.

Estos datos se someten a un preprocesamiento mediante los modelos generados a partir del conjunto de entrenamiento, utilizándolos tanto para el escalado como para la imputación de datos faltantes. Posteriormente, estos datos preprocesados se emplean en los modelos previamente entrenados para diagnóstico/pronóstico.

Capítulo 4. Resultados y Limitaciones

En este capítulo se reportan los resultados obtenidos en la presente tesis. El capítulo está dividido en 2 secciones: la sección 4.1 corresponde al estudio de los modelos de aprendizaje automático para realizar el diagnóstico de COVID-19, y en la sección 4.2 para el pronóstico (predicción de enfermedad grave). Los resultados de las subsecciones 4.1.1, 4.1.2 y 4.1.3 corresponden al análisis de los datos utilizando los datos de entrenamiento, mientras que los de las subsecciones 4.1.4 y 4.2 a los resultados del modelo al ser probado con los datos de prueba (ciegos), que fueron preprocesados con los modelos generados con el conjunto de entrenamiento para el escalamiento y la imputación de datos faltantes.

4.1Diagnóstico de COVID-19

4.1.1 Características individuales

Cada uno de los 4 modelos de ML (máquinas de soportes vectoriales (SVM, por sus siglas en inglés), bosques aleatorios, clasificador gaussiano ingenuo y regresión logística) fue entrenado y probado mediante LOOCV (con el conjunto de entrenamiento) para cada uno de los metabolitos posibles. La tabla 1 Muestra la exactitud balanceada en cada modelo para los 5 metabolitos que alcanzaron un mejor desempeño. Todos estos modelos fueron capaces de superar el 0.80 de exactitud balanceada para el diagnóstico, utilizando una única característica. Sin embargo, los resultados aún están por debajo del 0.90, y por lo tanto por debajo del rendimiento que comúnmente se reportan utilizando imágenes médicas.

Tabla 1. Exactitud balanceada para los 5 modelos que alcanzaron un mayor rendimiento en alguno de los 4 algoritmos de ML probados.

	SVM	Bosques Aleatorios	Regresión Logística	Bayesiano Ingenuo
LysoPC_a_C18:2	0.845	0.836	**0.868**	0.868
kynurenina/triptófano	**0.866**	0.856	0.76	0.866
Fenilalanina	0.538	0.812	0.812	**0.818**
kynurenina	0.528	0.747	0.813	**0.813**
LysoPC_a_C18:1	0.66	**0.809**	0.794	0.794

Para cada uno de estos modelos que utilizan metabolitos como características individuales se muestra la matriz de confusión del algoritmo que logró el mejor desempeño en la figura 16, y en la tabla 2 se muestra el rendimiento de estos en cada una de las 6 métricas evaluadas.

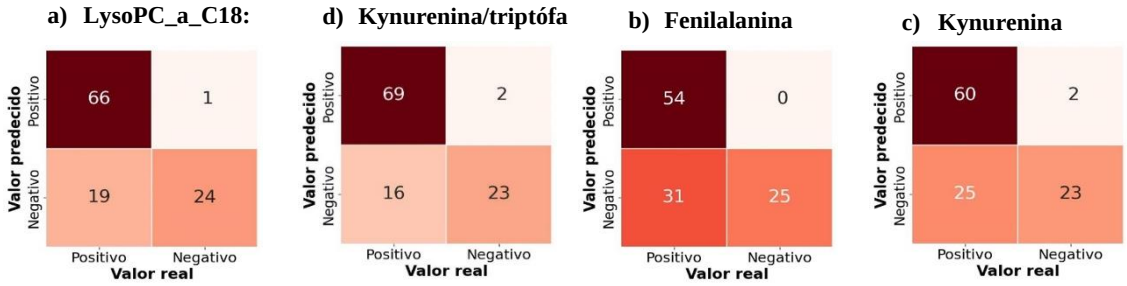

Figura 16. Matriz de confusión del mejor modelo con características individuales para cada uno de los metabolitos con mayor rendimiento.

Tabla 2. Rendimiento de los modelos con características individuales para los metabolitos con exactitud balanceada mayor a 0.80.

	Exactitud balanceada	F1	Precisión	Sensibilidad	Especificidad	ROC AUC
LysoPC_a_C18:2	0.868	0.868	0.985	0.776	0.960	0.900
kynurenina/triptófano	0.866	0.885	0.972	0.812	0.920	0.867
Fenilalanina	0.818	0.777	1.000	0.635	1.000	0.776
kynurenina	0.813	0.816	0.968	0.706	0.920	0.819

Estos modelos obtienen un rendimiento muy similar entre ellos. La predicción realizada por estos modelos es muy confiable cuando es positiva (alta precisión), y además clasifica casi perfectamente a los pacientes sanos (alta especificidad). Sin embargo, los modelos tienen un alto número de falsos negativos, lo que implica que hay una gran cantidad de personas enfermas que no son clasificadas correctamente (baja sensibilidad).

Así, aunque los resultados muestran alto rendimiento en precisión y especificidad, es importante encontrar un modelo con una mayor capacidad de clasificar correctamente a los pacientes enfermos.

4.1.2 Selección de características

La figura 17 muestra los metabolitos seleccionados por cada una de las 4 técnicas de selección de características empleadas, ordenados (de izquierda a derecha) de acuerdo al número de conjuntos en los que fueron considerados.

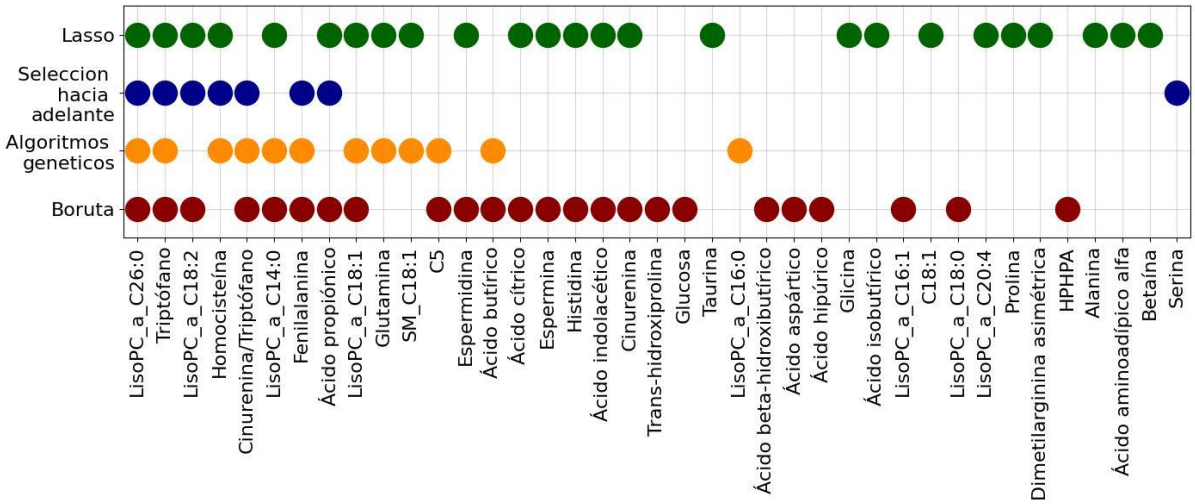

Figura 17. Metabolitos seleccionados por las 4 técnicas de selección de características utilizados como propuestas para la generación del modelo de Aprendizaje Automático para el diagnóstico del COVID-19.

En total, se identificaron 39 metabolitos a través de al menos alguna de las técnicas empleadas, excluyendo así los 67 restantes del conjunto original de características. Entre estos, resaltan 2 metabolitos que fueron seleccionados por las 4 técnicas: LysoPC_a_C26:0 y triptófano, seguido de otros 7 que fueron seleccionados en 3 de las 4 técnicas analizadas: LysoPC_a_C18:2, homocisteína, kynurenina/triptófano, LysoPC_a_C14:0, Fenilalanina, ácido propiónico y LysoPC_a_C18:1. Estos 9 metabolitos podrían considerarse como los más significativos y proporcionar información relevante. Además, 5 metabolitos fueron seleccionados por 2 de las 4 técnicas, mientras que los restantes fueron elegidos únicamente por uno de los métodos.

De los cuatro conjuntos de metabolitos propuestos, el que se obtuvo mediante la técnica de lasso es el más reducido, con solo 8 metabolitos, y por lo tanto el más práctico, seguido del encontrado mediante algoritmos genéticos con 12. En contraste, los conjuntos con más características son los obtenidos mediante las técnicas de boruta y lasso, con 24 y 25, respectivamente.

4.1.3 Selección del modelo

Cada uno de los conjuntos de metabolitos propuestos anteriormente fue aprobado por 4 diferentes algoritmos de ML para clasificación: máquinas de soportes vectoriales (SVM, por sus siglas en inglés), bosques aleatorios, clasificador gaussiano ingenuo y regresión logística, por lo que en total se probaron 16 modelos con las distintas combinaciones, todos ellos mediante validación cruzada dejando uno afuera. Además, en cada uno de ellos se realizó una búsqueda de hiperparámetros mediante una búsqueda de cuadrícula (con los parámetros mostrados en el apéndice D).

La figura 18 muestra los valores obtenidos por cada uno de los 16 modelos en exactitud balanceada. Así mismo, las tablas 3, 4, 5, 6, 7 y 8 muestran los rendimientos de estos mismos modelos en cada una de las métricas evaluadas (exactitud balanceada, F1, ROC AUC, sensibilidad, precisión y especificidad), los cuales se pueden verse en la en los diagramas de radar de la figura 19.

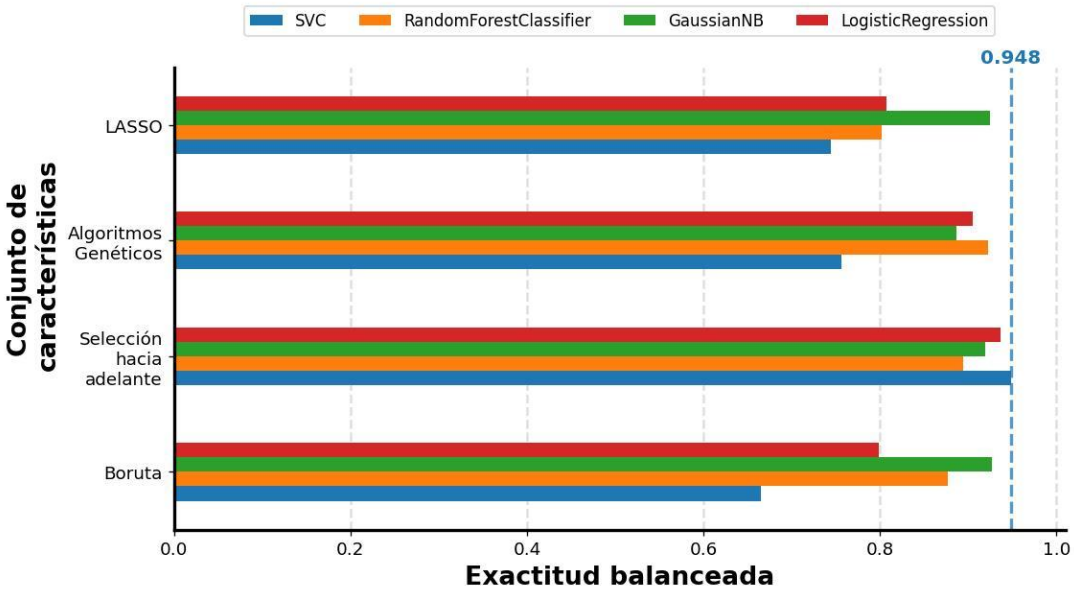

Figura 18. Exactitud balanceada para cada uno de los 16 distintos modelos creados con la combinación de los 4 conjuntos de características y los 4 algoritmos de ML, cada uno después de realizar búsqueda de hiperparámetros mediante una cuadrícula.

Tabla 3. Exactitud balanceada de los 16 modelos generados con los 4 algoritmos de ML y los 4 conjuntos de metabolitos obtenidos en la selección de características, cada uno después de realizar búsqueda de hiperparámetros mediante una cuadrícula. Marcado con azul el modelo propuesto.

	Boruta	**Selección hacia adelante**	**Algoritmos genéticos**	**Lasso**
SVC	0.665	**0.948**	0.756	0.745
Bosques Aleatorios	0.876	0.894	**0.922**	0.802
Bayesiano Ingenuo	**0.927**	0.919	0.887	**0.925**
Regresión Logística	0.799	0.936	0.905	0.807

Tabla 4. F1 de los 16 modelos generados con los 4 algoritmos de ML y los 4 conjuntos de metabolitos obtenidos en la selección de características, cada uno después de realizar búsqueda de hiperparámetros mediante una cuadrícula. Marcado con azul el modelo propuesto.

	Boruta	**Selección hacia adelante**	**Algoritmos genéticos**	**Lasso**
SVC	0.856	0.958	0.951	0.885
Bosques Aleatorios	**0.972**	0.927	0.927	**0.967**
Bayesiano Ingenuo	0.964	0.975	**0.954**	0.927
Regresión Logística	0.942	**0.988**	0.945	0.890

Tabla 5. Área bajo la curva ROC de los 16 modelos generados con los 4 algoritmos de ML y los 4 conjuntos de metabolitos obtenidos en la selección de características, cada uno después de realizar búsqueda de hiperparámetros mediante una cuadrícula. Marcado con azul el modelo propuesto.

	Boruta	**Selección hacia adelante**	**Algoritmos genéticos**	**Lasso**
SVC	0.856	0.961	0.951	0.885
Bosques Aleatorios	**0.972**	**0.993**	**0.969**	0.967
Bayesiano Ingenuo	0.964	0.975	0.954	**0.976**
Regresión Logística	0.942	0.988	0.945	0.890

Tabla 6. Sensibilidad de los 16 modelos generados con los 4 algoritmos de ML y los 4 conjuntos de metabolitos obtenidos en la selección de características, cada uno después de realizar búsqueda de hiperparámetros mediante una cuadrícula. Marcado con azul el modelo propuesto.

	Boruta	**Selección hacia adelante**	**Algoritmos genéticos**	**Lasso**
SVC	0.929	0.929	0.953	0.929
Bosques Aleatorios	**0.953**	**0.988**	**0.965**	**0.965**
Bayesiano Ingenuo	0.894	0.918	0.894	0.929
Regresión Logística	0.918	0.953	0.929	0.894

Tabla 7. Precisión de los 16 modelos generados con los 4 algoritmos de ML y los 4 conjuntos de metabolitos obtenidos en la selección de características, cada uno después de realizar búsqueda de hiperparámetros mediante una cuadrícula. Marcado con azul el modelo propuesto.

	Boruta	**Selección hacia adelante**	**Algoritmos genéticos**	**Lasso**
SVC	0.840	0.987	0.880	0.878
Bosques Aleatorios	0.942	0.944	**0.965**	0.901
Bayesiano Ingenuo	**0.987**	0.975	0.962	**0.975**
Regresión Logística	0.907	**0.976**	0.963	0.916

Tabla 8. Especificidad de los 16 modelos generados con los 4 algoritmos de ML y los 4 conjuntos de metabolitos obtenidos en la selección de características, cada uno después de realizar búsqueda de hiperparámetros mediante una cuadrícula.

	Boruta	**Selección hacia adelante**	**Algoritmos genéticos**	**Lasso**
SVC	0.400	**0.960**	0.560	0.560
Bosques Aleatorios	0.800	0.800	**0.880**	0.640
Bayesiano Ingenuo	**0.960**	**0.920**	**0.880**	**0.920**
Regresión Logística	0.680	**0.920**	**0.880**	0.720

Para cada uno de los conjuntos de metabolitos seleccionados es posible obtener una exactitud balanceada mayor al 0.90 en alguno de los posibles algoritmos: el conjunto obtenido por boruta alcanza una exactitud balanceada de 0.927 utilizando el clasificador bayesiano ingenuo, el obtenido por selección hacia adelante alcanza un 0.948 con SVC, el obtenido por algoritmos genéticos alcanza 0.922 utilizando bosques aleatorios, y el obtenido por lasso un 0.925 utilizando el clasificador bayesiano ingenuo. Aunque no existe una gran diferencia entre estos cuatro modelos (respecto a la exactitud balanceada), el mejor es el obtenido utilizando el conjunto de características obtenido por selección hacia adelante con el algoritmo de SVC, el cual cuenta con una menor la cantidad de características.

En general los modelos obtienen un alto resultado tanto en F1 como en el área bajo la curva ROC, con valores cercanos a 1 (con una media de 0.950 en ambas métricas), indicando que tienden a ser clasificadores confiables. Así mismo, todos los modelos obtienen un alto desempeño tanto en sensibilidad (capacidad de clasificar correctamente a los pacientes enfermos) como en precisión (probabilidad de que una clasificación positiva haya sido correcta), con una media de 0.926 entre los 16 modelos. En cambio, los valores más bajos se obtienen en la especificidad (capacidad de clasificar correctamente a los pacientes sanos) con valores que van desde 0.40 hasta 0.96, y una media de 0.777; sin embargo, esto es en parte debido a la baja cantidad de pacientes sanos en el conjunto de entrenamiento, con respecto al número de pacientes enfermos. En resumen, los modelos tienen una alta capacidad para detectar a los pacientes enfermos y una menor capacidad de detectar a los pacientes sanos.

A pesar de que todos los conjuntos de características parecen prometedores, en general, se propone el modelo generado por el algoritmo de SVC con el conjunto de características generado por selección hacia adelante, que cuenta con únicamente 8 metabolitos, para su implementación y uso.

La figura 20 muestra la matriz de confusión obtenida por este modelo (mediante validación cruzada dejando uno afuera). En la figura 20a, se muestra la matriz de confusión para diferenciar entre los casos positivos y negativos, mientras que la figura 20b muestra además para cada uno de los cuatro grupos (G1 los negativos, y G2 + G3 + G4 los positivos).

Este modelo genera únicamente 1 falsos positivos y 6 falsos negativos, clasificando correctamente a 103 de los 110 pacientes. Para los falsos positivos (figura 20b), 5 de ellos corresponden a

pacientes con enfermedad leve o moderada, y uno de ellos a un paciente grave.

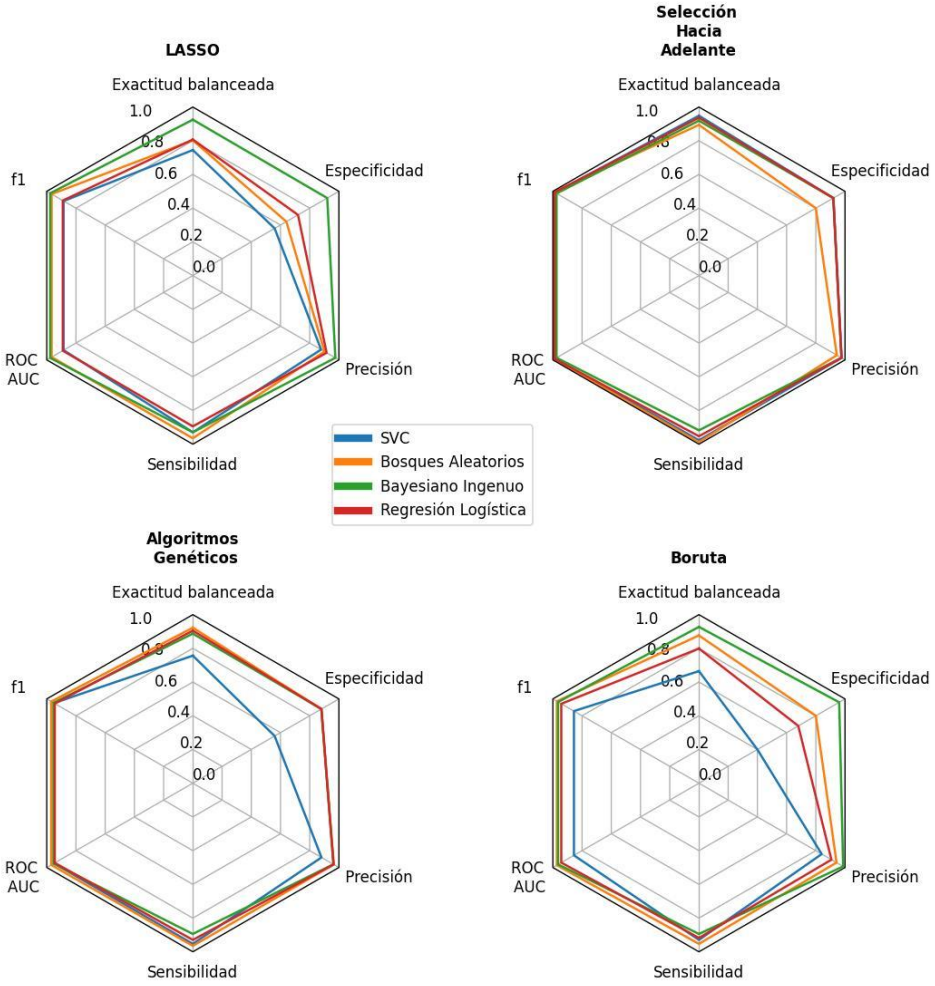

Figura 19. Rendimiento en 6 métricas evaluadas de cada uno de los 16 distintos modelos creados con la combinación de los 4 conjuntos de características y los 4 algoritmos de ML, cada

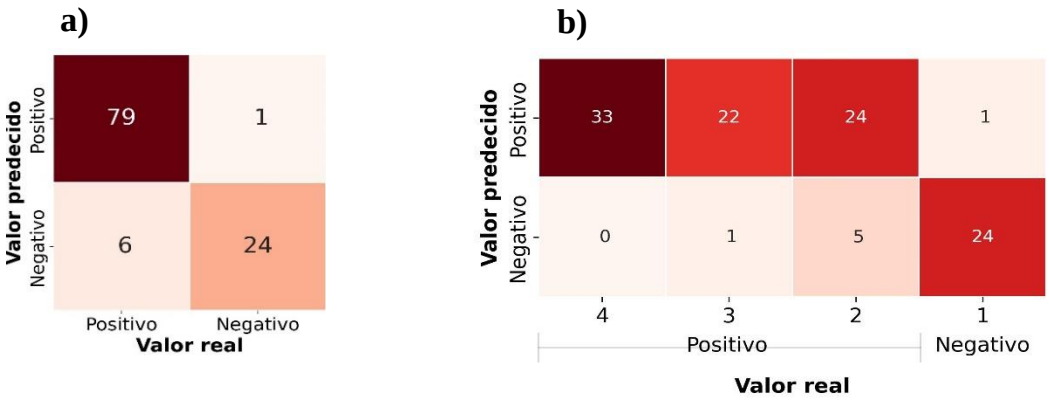

Figura 20. a) matriz de confusión del modelo propuesto para el diagnóstico de COVID-19 obtenida mediante LOOCV en el conjunto de prueba y b) la misma matriz de confusión, uno después de realizar búsqueda de hiperparámetros mediante una cuadrícula.

La tabla 9 muestra el rendimiento obtenido por este modelo para cada una de las métricas evaluadas. En general, todos estos valores se encuentran por encima de 0.90, e incluso cercanos a 1, por lo que en general se puede considerar como un alto rendimiento, capaz de clasificar a los pacientes con una alta confiabilidad, dentro del rango de los valores reportados para las técnicas utilizadas actualmente.

Tabla 9. Rendimiento del modelo seleccionado para el diagnóstico de COVID-19 obtenido mediante LOOCV en el conjunto de entrenamiento en las diferentes métricas evaluadas.

Métrica	**Valor obtenido**
Exactitud balanceada	0.958
F1	0.957
ROC AUC	0.962
Sensibilidad	0.929
Especificidad	0.960
Precisión	0.988

En la figura 21 se muestra la distribución de las predicciones realizadas por el modelo tanto para los pacientes enfermos (positivos) como para los sanos (negativos). La distribución de las

predicciones para los pacientes positivos tiene un sesgo positivo, con una media de 0.89, mientras que la de los pacientes sanos tiene un sesgo negativo con media de 0.23, ambas con una desviación estándar similar de 0.22, por lo que en general las predicciones del modelo tienden a ser muy confiables.

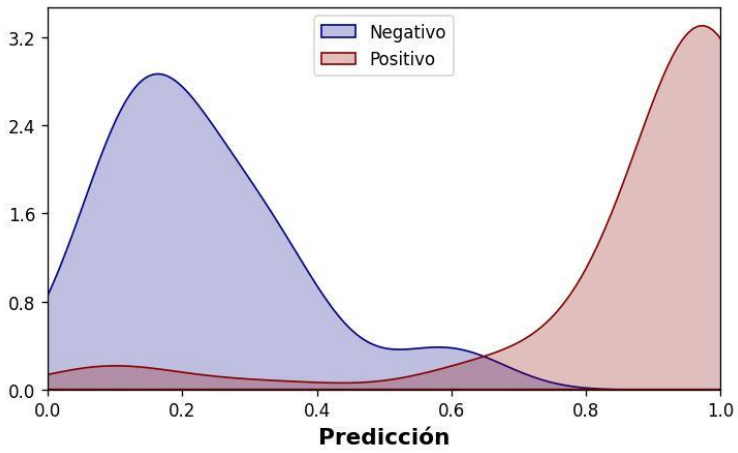

Figura 21. Distribución de las predicciones del modelo para los pacientes sanos (negativos) y enfermos (positivos) en el conjunto de entrenamiento, mediante LOOCV.

4.1.4 Prueba ciega

Para la prueba ciega se utilizó el conjunto de prueba separado inicialmente que consiste de 48 pacientes (12 pacientes sanos y 36 enfermos, 12 por cada uno de los grupos G1, G2, G3 y G4), utilizando el modelo propuesto (SVC utilizando el conjunto de características obtenido por selección hacia adelante) entrenado por todos los datos del conjunto de entrenamiento.

La tabla 10 muestra el rendimiento de este modelo en las diferentes métricas evaluadas, y en la figura 22 se muestra la matriz de confusión.

Tabla 10. Rendimiento en el conjunto de entrenamiento del mejor modelo generado para el diagnóstico de COVID-19.

Métrica	**Valor obtenido**
Exactitud balanceada	0.917
F1	0.917
ROC AUC	0.822
Sensibilidad	0.917
Especificidad	0.750
Precisión	0.917

Aunque el rendimiento del modelo baja con respecto al obtenido en el conjunto de entrenamiento utilizando LOOCV, obteniendo una exactitud de 0.833, debido a la especificidad obtenida de 0.750, los resultados del modelo son similares a los obtenidos en el conjunto de entrenamiento, con una alta capacidad de diagnosticar a los casos positivos (alta sensibilidad y precisión), y una menor capacidad para clasificar a los casos negativos (baja especificidad).

El modelo genera tanto 3 falsos positivos como 3 falsos negativos, y de estos últimos, 2 corresponden a pacientes con enfermad leve o moderada, y uno de ellos a pacientes muy graves. Al igual que en el conjunto de entrenamiento, la mayoría de los falsos negativos corresponden a pacientes con enfermedad leve a moderada (grupo G2).

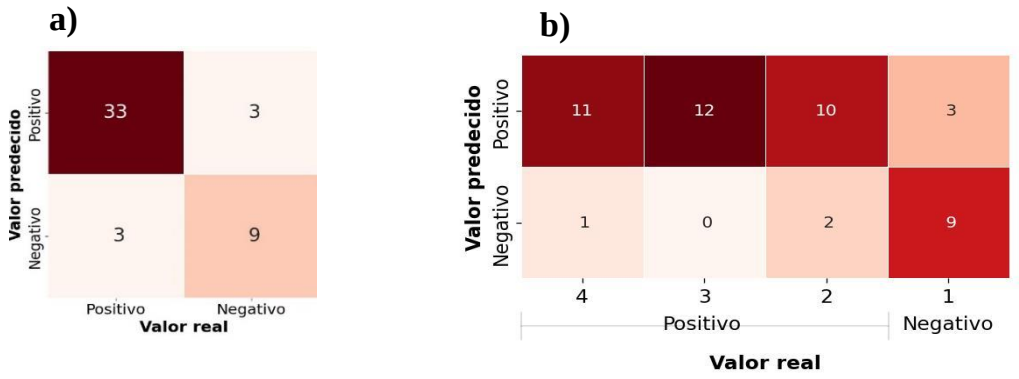

Figura 22. a) matriz de confusión del modelo propuesto para el diagnóstico de COVID-19 obtenida en la prueba ciega y b) la misma matriz de confusión, detallada para el grupo positivo real (G2, G3 y G4).

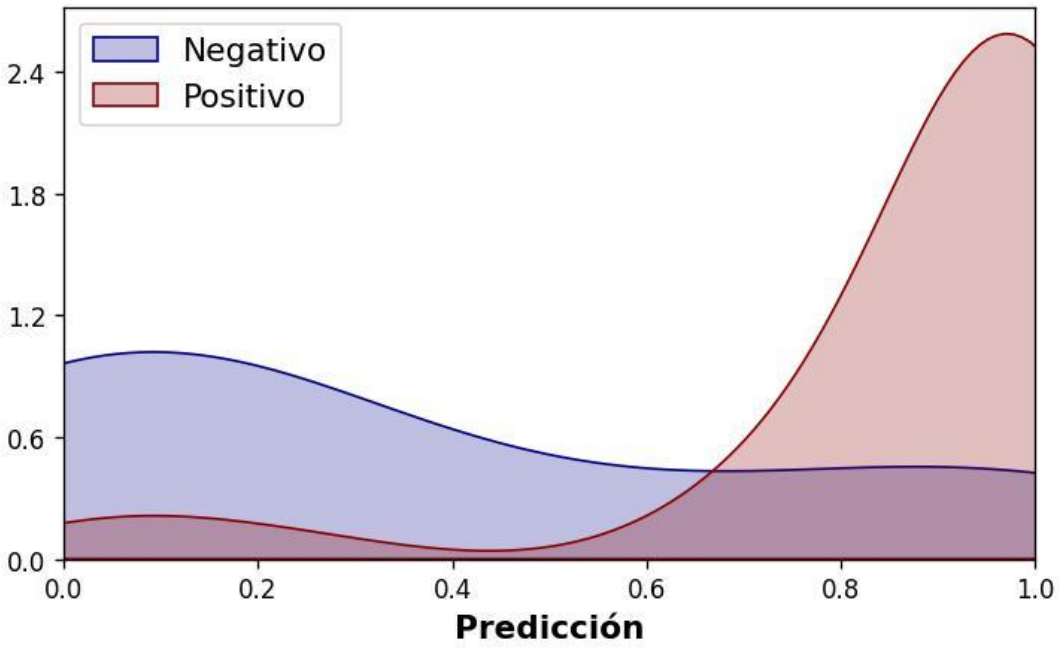

Figura 23. Distribución de las predicciones del modelo para los pacientes sanos (negativos) y enfermos (positivos) en el conjunto de prueba.

La figura 23 muestra la distribución que tienen tanto las predicciones positivas como las negativas. En este caso, las predicciones realizadas para los pacientes enfermos siguen estando cercanas a 1 (con una media de 0.88 y una desviación estándar de 0.25), mientras que las realizadas para los pacientes sanos están más cerca del 0 con una media de 0.35, pero se encuentran más dispersas (con una desviación estándar de 0.38), debido a que el modelo es capaz de clasificar correctamente a los pacientes enfermos con una alta confiabilidad, pero tiene un menor rendimiento clasificando a los pacientes sanos. Sin embargo, estos resultados son menos estables estadísticamente debido al bajo número de pacientes enfermos utilizados.

En la figura 24 se muestra también la curva ROC para los resultados del modelo tanto en el conjunto de prueba (obtenidos mediante LOOCV) como en el conjunto de entrenamiento. Para el conjunto de prueba, la curva ROC muestra casi un clasificador perfecto con un área bajo la curva de 0.95, mientras que la curva ROC para el conjunto de prueba se encuentra un poco por debajo (debido a la disminución del rendimiento del clasificador), con un área bajo la curva de 0.822.

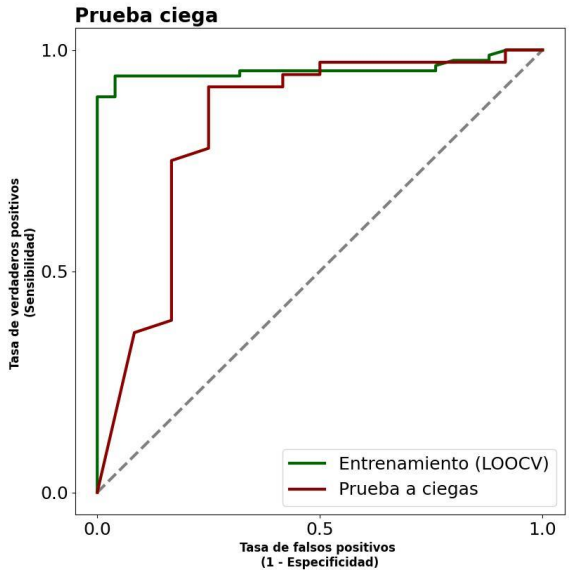

Figura 24. Curva ROC del modelo para el diagnóstico de COVID-19 tanto en el conjunto de entrenamiento (mediante LOOCV) como en el conjunto de prueba.

4.2 Predicción de enfermedad grave

El mismo modelo anterior (generado por el algoritmo de SVC con el conjunto de metabolitos seleccionado mediante selección hacia adelante) es probado para realizar un pronóstico de enfermedad por COVID-19 (predicción de enfermedad grave), por lo que el modelo es entrenado en este caso para clasificar a los pacientes ambulatorios (con enfermedad leve o moderada, grupo G2) contra los pacientes graves o muy graves (grupos G3 y G4). El entrenamiento del modelo es realizado con el conjunto de entrenamiento, y es probado posteriormente con el conjunto de prueba.

La figura 25 muestra la matriz de confusión del modelo al clasificar al grupo de pacientes ambulatorios (G2) contra los pacientes con enfermedad grave (G3 y G4) , y en la tabla 11 se encuentran los resultados de la evaluación de este modelo en las diferentes métricas utilizadas. Similar al resultado del modelo para realizar diagnóstico de COVID-19, el modelo es especialmente bueno para clasificar a los pacientes que tendrán una enfermedad grave, con únicamente 3 falsos negativos y una sensibilidad de 0.88, pero con un rendimiento menor para

clasificar a los pacientes ambulatorios, generando un total de 5 falsos positivos y una especificidad de 0.58.

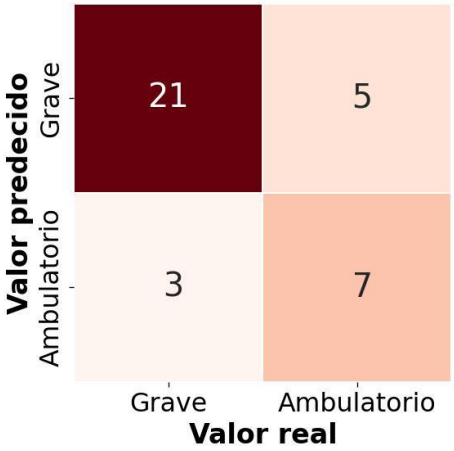

Figura 25. Matriz de confusión del modelo en el conjunto de prueba para predicción de enfermedad grave.

Tabla 11. Rendimiento del modelo en el conjunto de prueba para predicción de enfermedad grave en las diferentes métricas utilizadas.

Métrica	**Valor obtenido**
Exactitud balanceada	0.729
F1	0.840
ROC AUC	0.792
Sensibilidad	0.878
Especificidad	0.583
Precisión	0.808

En la figura 26 se muestra la distribución de las predicciones realizadas por el modelo para los pacientes con enfermedad leve o moderada (G2, ambulatorios) y graves (grupos G3 y G4). En este caso, las distribuciones tienen un alto grado de solapamiento debido a los falsos positivos y negativos, con una media de 0.46 y 0.68 para las clases negativas y positivas, respectivamente, y con una desviación estándar alrededor de 0.20 para ambas clases (0.19 para la clase positiva y 0.21 para la clase negativa). A pesar de que existe un amplio solapamiento entre ambas distribuciones, sí existe una diferencia estadísticamente significativa entre sus medias ($p<0.01$).

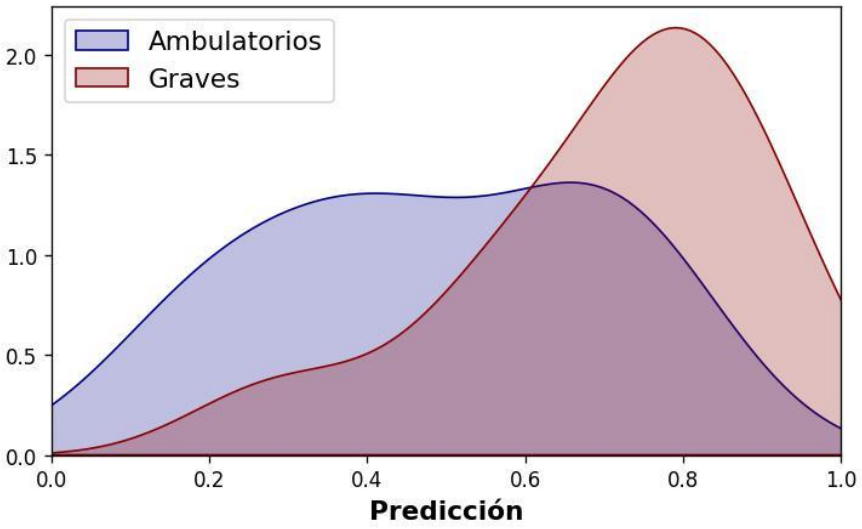

Figura 26. Distribución de las predicciones del modelo para los pacientes con enfermedad leve o moderada (negativos) y pacientes con enfermedad grave o muy grave (positivos) en el conjunto de prueba.

En la figura 27 se muestra la curva ROC de este modelo, el cual alcanza un área bajo la curva de 0.792.

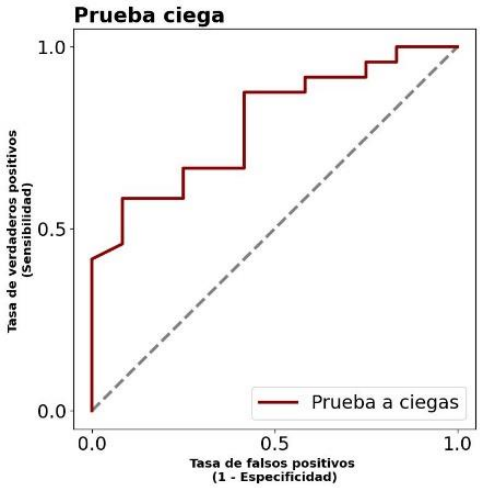

Figura 27. Curva ROC del modelo para clasificación de pacientes enfermos leves o moderados con pacientes enfermos graves o muy graves para el conjunto de prueba.

Capítulo 5. Discusión y conclusiones

En este capítulo se describen las conclusiones del trabajo. Comenzando por la sección 5.1, se describe la discusión de los resultados, y posteriormente las conclusiones en la sección 5.2. La sección 5.3 describe los objetivos alcanzados y en la sección 5.4 se describen los resultados en comparación con la hipótesis con la que parte la investigación. Finalmente, en la sección 5.5 se describen cuáles son las contribuciones de este trabajo de investigación.

5.1 Discusión

Debido a la necesidad de un consenso sobre cuáles son los metabolitos específicos, y en qué manera deben ser utilizados para estos fines, en este trabajo se realizó una selección de metabolitos con el fin de ser utilizados como características de un modelo de aprendizaje automático para el diagnóstico y pronóstico de pacientes con dicha enfermedad, utilizando 4 técnicas distintas de aprendizaje automático. Estos resultados concuerdan en gran medida con las propuestas anteriores, ya que la mayoría de estos metabolitos seleccionados están relacionados con el metabolismo de los lípidos (como el LysoPC_a_C26:0, LysoPC_a_C18:2, LysoPC_a_C14:0, entre otros), la vía triptófano/quinurenina (como el triptófano, el radio quinurenina/triptófano, la quinurenina y el ácido indolacético), el metabolismo energético (como el ácido propiónico, el ácido butírico, el ácido beta-hidroxibutírico, el ácido cítrico y la glucosa),.

A pesar de que se propuso el modelo SVC utilizando el conjunto generado por selección hacia adelante como el mejor modelo, con cada uno de los conjuntos se obtuvo un rendimiento mayor al 0.90 en exactitud balanceada con alguno de los 4 posibles modelos utilizados, y similarmente, con cada uno de los 4 modelos se logró también alcanzar un rendimiento similar utilizando alguno de los posibles conjuntos, por lo que en general podría haber más de un modelo posible con un rendimiento similar.

Aunque los resultados en el conjunto de entrenamiento fueron muy altos y prometedores, comparables con las técnicas actuales utilizadas para el diagnóstico de la enfermedad, durante la prueba estos rendimientos fueron considerablemente más bajos, aunque aún dentro del rango de

las pruebas RT-PCR y del uso de imágenes médicas. Sin embargo, esto indica que los modelos fueron sobre ajustados al conjunto de entrenamiento, más probablemente debido al bajo número de datos en el conjunto de entrenamiento. En una implementación con un conjunto de datos mayor, es probable que los rendimientos en la prueba ciega se acerquen más a los obtenidos en el conjunto de entrenamiento. Por otro lado, para el pronóstico de los pacientes (clasificación de pacientes ambulatorios vs graves) el modelo no obtuvo un rendimiento igual de alto que en el caso anterior. Por un lado, esto último fue debido a que la metodología fue desarrollada para la primera tarea, y por otro, debido a que el desenlace de la enfermedad de los pacientes depende en gran medida de los tratamientos y las medidas tomadas durante su progreso, por lo que las condiciones de los pacientes son muy distintas entre ellos.

Aunque todavía existe una validación médica para un futuro uso clínico, tanto el uso de modelos de aprendizaje simples e interpretables, como la fundamentación de los resultados con estudios previos y futuros podría facilitar este proceso.

5.2 Conclusiones

Este trabajo de investigación estuvo centrado en el desarrollo de un modelo de aprendizaje automático para su uso como herramienta de diagnóstico y pronóstico de COVID-19.

Gracias al uso de diferentes técnicas de selección de características, se lograron seleccionar 4 conjuntos distintos de metabolitos que mostraron un alto rendimiento para el diagnóstico de la enfermedad con alguno de los 4 modelos de aprendizaje automático durante el entrenamiento.

El mejor modelo para el diagnóstico de la enfermedad fue el de SVC con un total de 8 metabolitos como características (lysoPC_a_26:0, triptófano, lysoPC_a_18:2, homocisteína, quinurenina/triptófano, fenilalanina, ácido propiónico y serina) que obtuvo un 94.8% de exactitud balanceada durante el entrenamiento y un 83.3% en la prueba ciega, los cuales se encuentran dentro del rango de otras técnicas utilizadas actualmente como las pruebas RT-PCR y el uso de imágenes médicas, por lo que junto con el apoyo de un especialista, podría ser utilizado como una alternativa a las técnicas existentes. En particular, estos modelos mostraron tener una alta eficiencia detectando los casos positivos (con 91.2% tanto en sensibilidad como en precisión durante la

prueba ciega). Muy probablemente la caída en el rendimiento fue debido a un sobre ajuste del modelo ocasionado por la baja cantidad de datos de entrenamiento, por lo que un aumento en los datos para una implementación de esta propuesta podría acercarse al rendimiento obtenido durante el entrenamiento.

Aunque estos resultados aún requieren de una mayor investigación, gracias al uso de modelos simples e interpretables, así como los resultados de otras investigaciones previas que han demostrado que el SARS-CoV-2 causa alteraciones médicas en los metabolitos y las rutas metabólicas encontradas en este trabajo, puede ayudar a una posible validación médica en un futuro próximo.

5.3 Implicaciones de la investigación

Los resultados de este trabajo muestran la posibilidad de emplear los niveles de metabolitos junto con algoritmos de ML para el diagnóstico de COVID-19, con una eficacia similar a las técnicas actuales, destacando su un alto desempeño para detectar los casos positivos. Dado que las técnicas utilizadas actualmente, como las pruebas de laboratorio e imágenes médicas, suelen ser costosas, invasivas y lentas, además de tener una disponibilidad limitada, especialmente en épocas de brotes masivos, estos modelos ofrecen una alternativa para el diagnóstico de COVID-19, así como la capacidad de poder ser adaptadas a nuevas variantes de los coronavirus.

Una implementación de estos modelos podría tener varias ventajas, como la alta precisión y eficacia para realizar un diagnóstico rápido, que son técnicas que no requieren procedimientos invasivos, suelen ser más rentables que otras técnicas de laboratorio, y permiten una detección temprana, antes de que la enfermedad avance hacia una etapa más crítica. Además, estos modelos ofrecen también la capacidad de hacer pronóstico de la enfermedad, lo que permitiría un mejor tratamiento de los pacientes enfermos y una mayor administración de los recursos.

Sin embargo, estos modelos requieren aún de una validación en ensayos clínicos antes de poderse utilizar de manera práctica, y es necesaria una recolección mayor de datos para su implementación.

5.4 Objetivos alcanzados

Todos los objetivos particulares de este trabajo de investigación fueron alcanzados, lo que permitió alcanzar el objetivo general de la investigación de generar y evaluar modelos de

aprendizaje automático para su uso como herramienta de diagnóstico y pronóstico de COVID-19 utilizando metabolitos, y proponer un modelo y arquitectura particular como una posible herramienta alternativa a las técnicas existentes.

5.5 Hipótesis demostradas

Los resultados de este trabajo concuerdan con la hipótesis de partida. Es posible utilizar modelos simples de aprendizaje automático con metabolitos como características para desarrollar un modelo capaz de obtener rendimientos para el diagnóstico y pronóstico dentro del rango de las técnicas utilizadas actualmente, y estos resultados concuerdan con los obtenidos en estudios previos similares realizados por diferentes autores.

5.6 Contribuciones de la investigación

Se propuso un modelo de aprendizaje automático como posible alternativa para el diagnóstico del COVID-19 que además muestra rendimientos prometedores para realizar el pronóstico de la enfermedad que podrían mejorar el seguimiento de los pacientes enfermos, además de ventajas como el diagnóstico temprano, el aumento en la cantidad de centros capaces de diagnosticar la enfermedad, y reducción en los costos para esta tarea. En un futuro, estos resultados podrían ser más ampliamente estudiados y evaluados por el sector médico como herramientas alternativas a las existentes.

Además, se publicó el artículo titulado “Selección de metabolitos como características de un modelo de bosques aleatorios para el diagnóstico del COVID-19” (figura 28) en la revista *Research in computer Science* del Instituto Politécnico Nacional (Vol. 153), presentado en el Congreso Mexicano de Inteligencia Artificial (COMIA).

ISSN 1870-4069

Selección de metabolitos como características de un modelo de bosques aleatorios para el diagnóstico del COVID-19

Hugo Alexis Torres-Pasillas, José María Celaya-Padilla,
Yamilé López-Hernández, Carlos Erick Galván-Tejada,
Alejandra García-Hernández, Pedro Daniel Alaniz-Lumbreras,
José Alejandro Morgan-Benita

Universidad Autónoma de Zacatecas,
Unidad Académica de Ingeniería Eléctrica,
México

ylopezher@conacyt.mx {hugo.tpasillas, jose.celaya,
ericgalvan, alegarcia, dalaniz, alejandro.morgan}@uaz.edu.mx

Resumen. El COVID-19 es una enfermedad reciente que surgió a finales de 2019 causado por un nuevo tipo de coronavirus. A pesar de los avances en la investigación del virus y el desarrollo tanto de vacunas como de posibles tratamientos, el diagnóstico de la enfermedad, especialmente de forma temprana, continúa siendo una de las mejores herramientas para combatir la enfermedad y su transmisión. El objetivo de este estudio es seleccionar el mejor conjunto de metabolitos como potenciales biomarcadores para el diagnóstico, que son utilizados como características de un modelo de bosques aleatorios. Para ello, se utilizaron 4 diferentes técnicas de selección de características que son utilizadas con frecuencia dentro del Aprendizaje Automático, y un conjunto de datos que contiene mediciones de 110 metabolitos de 158 pacientes sospechosos de COVID-19 (121 enfermos y 37 sanos confirmados por pruebas rt-PCR). Los resultados muestran cuatro distintos conjuntos de metabolitos capaces de diagnosticar el COVID-19 con un alto desempeño en 6 distintas métricas utilizadas. El conjunto con mejor rendimiento en el conjunto de entrenamiento consta de 15 metabolitos y logra tener un desempeño alto en la validación a ciegas (f1=0.921, exactitud balanceada=0.875, AUC=0.910), mientras que el conjunto con menor número de características (5) obtiene el segundo mejor rendimiento en el conjunto de entrenamiento pero el mejor desempeño en la validación a ciegas (f1=0.931, exactitud balanceada=0.896, AUC=0.858).

Palabras clave: COVID-19, aprendizaje automático, metabolitos, selección de características, diagnóstico.

Selection of Metabolites as Features of a Random Forest Model for COVID-19 Diagnosis

Figura 28. Artículo de investigación publicado en la revista Research in computer Science del Instituto Politécnico Nacional (Vol. 153), presentado en el Congreso Mexicano de Inteligencia Artificial (COMIA) como resultado del presente trabajo de investigación.

Referencias

[1] Abdelaziz, O. S., & Waffa, Z. (2020). Neuropathogenic human coronaviruses: A review. *Reviews in Medical Virology*, *30*(5), e2118. https://doi.org/10.1002/rmv.2118

[2] Ali, I., & Alharbi, O. M. L. (2020). COVID-19: Disease, management, treatment, and social impact. *Science of The Total Environment*, *728*, 138861. https://doi.org/10.1016/j.scitotenv.2020.138861

[3] Baiges-Gaya, G., Iftimie, S., Castañé, H., Rodríguez-Tomàs, E., Jiménez-Franco, A., López-Azcona, A. F., Castro, A., Camps, J., & Joven, J. (2023). Combining Semi-Targeted Metabolomics and Machine Learning to Identify Metabolic Alterations in the Serum and Urine of Hospitalized Patients with COVID-19. *Biomolecules*, *13*(1), 163. https://doi.org/10.3390/biom13010163

[4] Bardanzellu, F., & Fanos, V. (2022). Metabolomics, Microbiomics, Machine learning during the COVID-19 pandemic. *Pediatric Allergy and Immunology*, *33*(S27), 86–88. https://doi.org/10.1111/pai.13640

[5] Blasco, H., Bessy, C., Plantier, L., Lefevre, A., Piver, E., Bernard, L., Marlet, J., Stefic, K., Benz-de Bretagne, I., Cannet, P., Lumbu, H., Morel, T., Boulard, P., Andres, C. R., Vourc'h, P., Hérault, O., Guillon, A., & Emond, P. (2020). The specific metabolome profiling of patients infected by SARS-COV-2 supports the key role of tryptophan-nicotinamide pathway and cytosine metabolism. *Scientific Reports*, *10*(1), 16824. https://doi.org/10.1038/s41598-020-73966-5

[6] Bourgin, M., Durand, S., & Kroemer, G. (2023). Diagnostic, Prognostic and Mechanistic Biomarkers of COVID-19 Identified by Mass Spectrometric Metabolomics. *Metabolites*, *13*(3), 342. https://doi.org/10.3390/metabo13030342

[7] Camps, J., Castañé, H., Rodríguez-Tomàs, E., Baiges-Gaya, G., Hernández-Aguilera, A., Arenas, M., Iftimie, S., & Joven, J. (2021). On the Role of Paraoxonase-1 and Chemokine Ligand 2 (C-C motif) in Metabolic Alterations Linked to Inflammation and Disease. A 2021 Update. *Biomolecules*, *11*(7), 971. https://doi.org/10.3390/biom11070971

[8] Ciotti, M., Ciccozzi, M., Terrinoni, A., Jiang, W.-C., Wang, C.-B., & Bernardini, S. (2020). The COVID-19 pandemic. *Critical Reviews in Clinical Laboratory Sciences*,

57(6), 365–388. https://doi.org/10.1080/10408363.2020.1783198

[9] Corman, V. M., Muth, D., Niemeyer, D., & Drosten, C. (2018). Hosts and Sources of Endemic Human Coronaviruses. En *Advances in Virus Research* (Vol. 100, pp. 163–188). Elsevier. https://doi.org/10.1016/bs.aivir.2018.01.001

[10] Cui, J., Li, F., & Shi, Z.-L. (2019). Origin and evolution of pathogenic coronaviruses. *Nature Reviews Microbiology*, *17*(3), 181–192. https://doi.org/10.1038/s41579-018-0118-9

[11] Danlos, F.-X., Grajeda-Iglesias, C., Durand, S., Sauvat, A., Roumier, M., Cantin, D., Colomba, E., Rohmer, J., Pommeret, F., Baciarello, G., Willekens, C., Vasse, M., Griscelli, F., Fahrner, J.-E., Goubet, A.-G., Dubuisson, A., Derosa, L., Nirmalathasan, N., Bredel, D., … Kroemer, G. (2021). Metabolomic analyses of COVID-19 patients unravel stage-dependent and prognostic biomarkers. *Cell Death & Disease*, *12*(3), 258. https://doi.org/10.1038/s41419-021-03540-y

[12] Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94

[13] Davis, E. L., Lucas, T. C. D., Borlase, A., Pollington, T. M., Abbott, S., Ayabina, D., Crellen, T., Hellewell, J., Pi, L., CMMID COVID-19 working group, Medley, G. F., Hollingsworth, T. D., & Klepac, P. (2020). *An imperfect tool: Contact tracing could provide valuable reductions in COVID-19 transmission if good adherence can be achieved and maintained* [Preprint]. Public and Global Health. https://doi.org/10.1101/2020.06.09.20124008

[14] Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, *26*(1), 51–78. https://doi.org/10.1002/mas.20108

[15] Drosten, C., Günther, S., Preiser, W., Van Der Werf, S., Brodt, H.-R., Becker, S., Rabenau, H., Panning, M., Kolesnikova, L., Fouchier, R. A. M., Berger, A., Burguière, A.-M., Cinatl, J., Eickmann, M., Escriou, N., Grywna, K., Kramme, S., Manuguerra, J.-C., Müller, S., … Doerr, H. W. (2003). Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome. *New England Journal of Medicine*, *348*(20), 1967–1976. https://doi.org/10.1056/NEJMoa030747

[16] Ghafouri-Fard, S., Mohammad-Rahimi, H., Motie, P., Minabi, M. A. S., Taheri, M., & Nateghinia, S. (2021). Application of machine learning in the prediction of COVID-19 daily new cases: A scoping review. *Heliyon*, *7*(10), e08143. https://doi.org/10.1016/j.heliyon.2021.e08143

[17] Hamre, D., & Procknow, J. J. (1966). A New Virus Isolated from the Human Respiratory Tract. *Experimental Biology and Medicine*, *121*(1), 190–193. https://doi.org/10.3181/00379727-121-30734

[18] Hasan, M. R., Suleiman, M., & Pérez-López, A. (2021). Metabolomics in the Diagnosis and Prognosis of COVID-19. *Frontiers in Genetics*, *12*, 721556. https://doi.org/10.3389/fgene.2021.721556

[19] Hemdan, E. E.-D., El-Shafai, W., & Sayed, A. (2022). CR19: A framework for preliminary detection of COVID-19 in cough audio signals using machine learning algorithms for automated medical diagnosis applications. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-022-03732-0

[20] Hernández Melo, C. A. (2021). *Estudio metabólico sobre el efecto circadiano en el uso de metotrexato (MTX) en los peces cebra (Danio rerio)*.

[21] Jamil, S., Mark, N., Carlos, G., Cruz, C. S. D., Gross, J. E., & Pasnick, S. (2020). Diagnosis and Management of COVID-19 Disease. *American Journal of Respiratory and Critical Care Medicine*, *201*(10), P19–P20. https://doi.org/10.1164/rccm.2020C1

[22] Li, P., Ikram, A., Peppelenbosch, M. P., Ma, Z., & Pan, Q. (2021). Systematically Mapping Clinical Features of Infections With Classical Endemic Human Coronaviruses. *Clinical Infectious Diseases*, *73*(3), 554–555. https://doi.org/10.1093/cid/ciaa1386

[23] Lin, D.-Y., Gu, Y., Wheeler, B., Young, H., Holloway, S., Sunny, S.-K., Moore, Z., & Zeng, D. (2022). Effectiveness of Covid-19 Vaccines over a 9-Month Period in North Carolina. *New England Journal of Medicine*, *386*(10), 933–941. https://doi.org/10.1056/NEJMoa2117128

[24] Liu, T., Siegel, E., & Shen, D. (2022). Deep Learning and Medical Image Analysis for COVID-19 Diagnosis and Prediction. *Annual Review of Biomedical Engineering*, *24*(1), 179–201. https://doi.org/10.1146/annurev-bioeng-110220-012203

[25] Liu, X., & Locasale, J. W. (2017). Metabolomics: A Primer. *Trends in Biochemical*

Sciences, *42*(4), 274–284. https://doi.org/10.1016/j.tibs.2017.01.004

[26] McIntosh, K., Dees, J. H., Becker, W. B., Kapikian, A. Z., & Chanock, R. M. (1967). Recovery in tracheal organ cultures of novel viruses from patients with respiratory disease. *Proceedings of the National Academy of Sciences*, *57*(4), 933–940. https://doi.org/10.1073/pnas.57.4.933

[27] Mishra, N. K., Singh, P., & Joshi, S. D. (2021). Automated detection of COVID-19 from CT scan using convolutional neural network. *Biocybernetics and Biomedical Engineering*, *41*(2), 572–588. https://doi.org/10.1016/j.bbe.2021.04.006

[28] Montani, D., Savale, L., Noel, N., Meyrignac, O., Colle, R., Gasnier, M., Corruble, E., Beurnier, A., Jutant, E.-M., Pham, T., Lecoq, A.-L., Papon, J.-F., Figueiredo, S., Harrois, A., Humbert, M., & Monnet, X. (2022). Post-acute COVID-19 syndrome. *European Respiratory Review*, *31*(163), 210185. https://doi.org/10.1183/16000617.0185-2021

[29] Moulaei, K., Shanbehzadeh, M., Mohammadi-Taghiabad, Z., & Kazemi-Arpanahi, H. (2022). Comparing machine learning algorithms for predicting COVID-19 mortality. *BMC Medical Informatics and Decision Making*, *22*(1), 2. https://doi.org/10.1186/s12911-021-01742-0

[30] Noy, O., Coster, D., Metzger, M., Atar, I., Shenhar-Tsarfaty, S., Berliner, S., Rahav, G., Rogowski, O., & Shamir, R. (2022). A machine learning model for predicting deterioration of COVID-19 inpatients. *Scientific Reports*, *12*(1), 2630. https://doi.org/10.1038/s41598-022-05822-7

[31] Páez-Franco, J. C., Torres-Ruiz, J., Sosa-Hernández, V. A., Cervantes-Díaz, R., Romero-Ramírez, S., Pérez-Fragoso, A., Meza-Sánchez, D. E., Germán-Acacio, J. M., Maravillas-Montero, J. L., Mejía-Domínguez, N. R., Ponce-de-León, A., Ulloa-Aguirre, A., Gómez-Martín, D., & Llorente, L. (2021). Metabolomics analysis reveals a modified amino acid metabolism that correlates with altered oxygen homeostasis in COVID-19 patients. *Scientific Reports*, *11*(1), 6350. https://doi.org/10.1038/s41598-021-85788-0

[32] Payne, S. (2017). Family Coronaviridae. En *Viruses* (pp. 149–158). Elsevier. https://doi.org/10.1016/B978-0-12-803109-4.00017-9

[33] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and

Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. (2011). *Scikit-learn: Machine Learning in {P}ython}* (Vol. 12, pp. 2825–2830). Jounal of Machine Learning Research.

[34] Pierce, J. D., Shen, Q., Cintron, S. A., & Hiebert, J. B. (2022). Post-COVID-19 Syndrome. *Nursing Research*, *71*(2), 164–174. https://doi.org/10.1097/NNR.0000000000000565

[35] Pourkarim, F., Pourtaghi-Anvarian, S., & Rezaee, H. (2022). Molnupiravir: A new candidate for COVID-19 treatment. *Pharmacology Research & Perspectives*, *10*(1). https://doi.org/10.1002/prp2.909

[36] Ryan, D., Newnham, E. D., Prenzler, P. D., & Gibson, P. R. (2015). Metabolomics as a tool for diagnosis and monitoring in coeliac disease. *Metabolomics*, *11*(4), 980–990. https://doi.org/10.1007/s11306-014-0752-9

[37] Saadatmand, S., Salimifard, K., Mohammadi, R., Marzban, M., & Naghibzadeh-Tahami, A. (2022). Predicting the necessity of oxygen therapy in the early stage of COVID-19 using machine learning. *Medical & Biological Engineering & Computing*, *60*(4), 957–968. https://doi.org/10.1007/s11517-022-02519-x

[38] SAMPIERI, H. (2016). *Fundamentos de metodologia de la investigacion*. MCGRAW-HILL.

[39] Santos-López, G., Cortés-Hernández, P., Vallejo-Ruiz, V., & Reyes-Leyva, J. (2021). SARS-CoV-2: Generalidades, origen y avances en el tratamiento. *Gaceta Médica de México*, *157*(1), 4792. https://doi.org/10.24875/GMM.20000505

[40] Shahin, O. R., Alshammari, H. H., Taloba, A. I., & El-Aziz, R. M. A. (2022). Machine Learning Approach for Autonomous Detection and Classification of COVID-19 Virus. *Computers and Electrical Engineering*, *101*, 108055. https://doi.org/10.1016/j.compeleceng.2022.108055

[41] Shen, B., Yi, X., Sun, Y., Bi, X., Du, J., Zhang, C., Quan, S., Zhang, F., Sun, R., Qian, L., Ge, W., Liu, W., Liang, S., Chen, H., Zhang, Y., Li, J., Xu, J., He, Z., Chen, B., … Guo, T. (2020). Proteomic and Metabolomic Characterization of COVID-19 Patient Sera. *Cell*, *182*(1), 59-72.e15. https://doi.org/10.1016/j.cell.2020.05.032

[42] Shereen, M. A., Khan, S., Kazmi, A., Bashir, N., & Siddique, R. (2020). COVID-19 infection: Emergence, transmission, and characteristics of human coronaviruses. *Journal*

of Advanced Research, *24*, 91–98. https://doi.org/10.1016/j.jare.2020.03.005

[43] Singh, D., & Yi, S. V. (2021). On the origin and evolution of SARS-CoV-2. *Experimental & Molecular Medicine*, *53*(4), 537–547. https://doi.org/10.1038/s12276-021-00604-z

[44] Subramani, R., Poudel, S., Smith, K. D., Estrada, A., & Lakshmanaswamy, R. (2022). Metabolomics of Breast Cancer: A Review. *Metabolites*, *12*(7), 643. https://doi.org/10.3390/metabo12070643

[45] Tiwari, S., Chanak, P., & Singh, S. K. (2023). A Review of the Machine Learning Algorithms for Covid-19 Case Analysis. *IEEE Transactions on Artificial Intelligence*, *4*(1), 44–59. https://doi.org/10.1109/TAI.2022.3142241

[46] Vandenberg, O., Martiny, D., Rochas, O., van Belkum, A., & Kozlakidis, Z. (2021). Considerations for diagnostic COVID-19 tests. *Nature Reviews Microbiology*, *19*(3), 171–183. https://doi.org/10.1038/s41579-020-00461-z

[47] Wen, W., Chen, C., Tang, J., Wang, C., Zhou, M., Cheng, Y., Zhou, X., Wu, Q., Zhang, X., Feng, Z., Wang, M., & Mao, Q. (2022). Efficacy and safety of three new oral antiviral treatment (molnupiravir, fluvoxamine and Paxlovid) for COVID-19 : a meta-analysis. *Annals of Medicine*, *54*(1), 516–523. https://doi.org/10.1080/07853890.2022.2034936

[48] Xiong, Y., Ma, Y., Ruan, L., Li, D., Lu, C., Huang, L., & the National Traditional Chinese Medicine Medical Team. (2022). Comparing different machine learning techniques for predicting COVID-19 severity. *Infectious Diseases of Poverty*, *11*(1), 19. https://doi.org/10.1186/s40249-022-00946-4

[49] Yu, F., Lau, L.-T., Fok, M., Lau, J. Y.-N., & Zhang, K. (2021). COVID-19 Delta variants—Current status and implications as of August 2021. *Precision Clinical Medicine*, *4*(4), 287–292. https://doi.org/10.1093/pcmedi/pbab024

[50] Zhang, C., Yao, J., Hu, G., & Sch鳩t, T. (2020). Applying Feature-Weighted Gradient Decent K-Nearest Neighbor to Select Promising Projects for Scientific Funding. *Computers, Materials & Continua*, *64*(3), 1741–1753. https://doi.org/10.32604/cmc.2020.010306

[51] Zhang, F. (2021). Application of machine learning in CT images and X-rays of COVID-19 pneumonia. *Medicine*, *100*(36), e26855. https://doi.org/10.1097/MD.0000000000026855

[52] Zumla, A., Dar, O., Kock, R., Muturi, M., Ntoumi, F., Kaleebu, P., Eusebio, M., Mfinanga, S., Bates, M., Mwaba, P., Ansumana, R., Khan, M., Alagaili, A. N., Cotten, M., Azhar, E. I., Maeurer, M., Ippolito, G., & Petersen, E. (2016). Taking forward a 'One Health' approach for turning the tide against the Middle East respiratory syndrome coronavirus and other zoonotic pathogens with epidemic potential. *International Journal of Infectious Diseases*, *47*, 5–9. https://doi.org/10.1016/j.ijid.2016.06.012

Anexos

A. Trabajos Publicados

- Torres-Pasillas, H. A., Celaya-Padilla, J. M., López-Hernández, Y., Galván-Tejada, C. E., Garcıa-Hernández, A., Alaniz-Lumbreras, P. D., & Morgan-Benita, J. A. Selección de metabolitos como características de un modelo de bosques aleatorios para el diagnóstico del COVID-19.

B. Parámetros de los algoritmos de selección de características

Tabla 12. Parámetros utilizados por el método de algoritmos genéticos.

Parámetro	Valor	Definición
cv	25	Numero de divisiones que se utilizados en la validación cruzada.
scoring	balanced_accuracy	Métrica de evaluación utilizada.
population	50	Número de individuos en la población inicial.
generations	200	Total de generaciones realizadas en el algoritmo evolutivo.
elitism	True	Si mantener las mejores soluciones para la siguiente generación intactas.
max_features	True	Máximo de características seleccionadas.
crossover_probability	0.3	Probabilidad de cruce entre dos individuos.
mutation_probability	0.7	Probabilidad de mutación de los individuos.

Tabla 13. Parámetros utilizados por el algoritmo de boruta.

Parámetro	**Valor**	**Definición**
n_estimators	500	Cantidad de estimadores en el ensamble.

max_iter	50	Máximo número de iteraciones.
alpha	0.01	Nivel en los cuales los p-valores corregidos serán rechazados.

Tabla 14. Parámetros utilizados por el algoritmo de LASSO.

Parámetro	**Valor**	**Definición**
cv	25	Numero de divisiones que se utilizados en la validación cruzada.
scoring	neg_mean_squared_error	Métrica de evaluación utilizada.
alpha	0.1 (encontrado por GridSearchCV)	Parámetro de regularización L1.