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

Abstract. COVID-19 is a recent disease that emerged in late 2019 caused by a new type of coronavirus. Despite advances in virus research and the development of both vaccines and potential treatments, early and accurate

diagnosis of the disease remains one of the best tools to combat the disease and its transmission. The aim of this study is to select the best set of metabolites as potential biomarkers for diagnosis, which are used as features of a random forest model. To achieve this, four different feature selection techniques that are frequently used in Machine Learning, and a dataset containing measurements of 110 metabolites from 158 suspected COVID-19 patients (121 confirmed patients and 37 confirmed healthy by rt-PCR tests) were used. The results show four different sets of metabolites capable of diagnosing COVID-19 with high performance in six different metrics used. The set with the best performance in the training set consists of 15 metabolites and achieves high performance in blind validation (f1=0.921, balanced accuracy=0.875, AUC=0.910), while the set with the smallest number of features (5) obtains the second best performance in the training set but the best performance in blind validation (f1=0.931, balanced accuracy=0.896, AUC=0.858).

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

1. Introducción

El Síndrome Respiratorio Agudo Severo 2 (SARS-CoV 2), es un tipo de coronavirus que surgió a finales de 2019 en la ciudad de Wuhan, en la providencia de Hubei, en China central, después de que un hospital notificó el ingreso de un paciente con neumonía grave e insuficiencia respiratoria el 26 de diciembre de 2019 [23].

La enfermedad causada por el SARS-CoV 2, denominada COVID-19, se caracteriza principalmente por síntomas que incluyen fiebre, tos seca y dificultad para respirar, además de síntomas menos comunes como vómito, diarrea, y dolor abdominal que aparecen dentro de los 2 a 14 días siguientes después del contagio [2].

Aunque la enfermedad se presenta mayormente sin síntomas o con síntomas de leves a moderados (en más del 80 % de los casos), en algunos casos estos pueden empeorar y requerir hospitalización y el uso de ventilación artificial, o hasta conducir a la muerte [13].

El día 11 de marzo de 2020, la Organización Mundial de la Salud (OMS) declaró la pandemia del COVID-19, y a marzo de 2023 continúa siendo una problemática actual, con una cifra de contagios que supera ya los 758 millones de individuos y que sigue aumentando a un ritmo de alrededor de 140 mil casos nuevos por día (tomado como el promedio de los últimos 7 días), con más de 6.8 millones de muertes acumuladas, de acuerdo con los datos reportados por la OMS [22].

Aunque ya se han estudiado y propuesto algunos medicamentos como el molnupiravir, la fluvoxamina y el paxlovid que reducen la mortalidad y la hospitalización en aproximadamente un 67 %, a día de hoy no existen tratamientos específicos contra esta enfermedad [19]. Por otro lado, se han desarrollado ya diferentes vacunas en contra de esta enfermedad que han mostrado buenos resultados, las cuales se han aplicado a nivel global desde finales de 2021, con eficiencias mayores al 90 % a 2 meses después de la primera dosis, y 60 % después de los 7 meses [16].

Sin embargo, estas se han encontrado con diferentes desafíos, como el acceso desigual a las diferentes vacunas y problemas de distribución, o la aparición de nuevas variantes del virus que reducen su eficiencia [3]. Por ello, el diagnóstico de la enfermedad sigue siendo una de las mejores herramientas, especialmente si se hace en fases tempranas del contagio.

Durante las últimas décadas, el área de la Inteligencia Artificial (IA), especialmente sus rama de Aprendizaje Automático (ML por sus siglas en inglés) y Aprendizaje Profundo (DL por sus siglas en inglés) han mostrado un gran potencial tanto en el área de la salud como en otras áreas, en aplicaciones como el diagnóstico de enfermedades, recomendación de tratamientos, predicciones de riesgo, entre otras [1], en muchas ocasiones mostrando desempeños incluso superiores a los métodos más clásicos o a los humanos [6].

Un enfoque reciente para el diagnóstico de enfermedades ha sido el uso de la metabolómica, el cual además puede ser utilizado para entender las interacciones del virus en el cuerpo a niveles moleculares [10]. Mediante este enfoque, se ha permitido encontrar biomarcadores para el diagnóstico y pronóstico de enfermedades como el Virus de la Inmunodeficiencia Uumana (VIH), hepatitis B y C, entre otros, así como a identificar rutas metabólicas que son alteradas debido a la presencia de los patógenos causantes de estas enfermedades [4].

En este trabajo estudiamos el uso de metabolitos como características para un modelo de Bosques Aleatorios (RF por sus siglas en inglés) para el diagnóstico de la enfermedad de COVID-19. Utilizando un conjunto de datos presentado por López Y., et al. [17], donde se cuantifican 110 metabolitos de 121 pacientes con diferentes niveles de severidad (agrupados de acuerdo a su nivel de severidad en 3 grupos: 2, 3 y 4) y 37 personas sanas (grupo 1).

Mediante el uso de las técnicas de selección de características de selección hacia adelante, Boruta, algoritmos genéticos y Regresión LASSO, se selecciona el conjunto de metabolitos con mejor desempeño en el conjunto de entrenamiento que consta de 110 individuos (obtenido por validación cruzada dejando uno fuera), y posteriormente se utiliza un conjunto de validación de 48 pacientes (12 por cada grupo).

El artículo está organizado de la siguiente manera: la sección 1 presenta una introducción al tema desarrollado, así como su justificación, y en la sección 2 presentamos los trabajos y resultados anteriores relacionados.

En la sección 3 introducimos los métodos utilizados tanto para la selección del conjunto de metabolitos como para la validación, y en la sección 4 presentamos los experimentos realizados y los resultados obtenidos. En la sección 5 presentamos las conclusiones del presente trabajo. Finalmente, la sección 6 corresponde a los agradecimientos.

2. Trabajos relacionados

Las primeras investigaciones del uso de ML y DL para el diagnóstico de COVID-19 se centraron en el uso de las imágenes médicas de Tórax analizadas por algoritmos de ML y DL.

En [20], por ejemplo, utilizan esta técnica y proponen un sistema de diagnóstico asistido por computadora con un alto radio de detección, proponiendo que el uso de ML para detectar y categorizar a pacientes de COVID-19 mediantes estas imágenes médicas es la mejor forma para diagnosticar el virus.

Además, otros estudios han utilizado técnicas de ML y DL para el diagnóstico de COVID-19 con alto desempeño, mediante el uso de características como síntomas [25], señales de audio [11], entre otras.

Por otra parte, 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 y virología de diagnóstico [10].

En [21], Bardanzellu F., et al. mencionan que, en el futuro próximo, la combinación de la metabolómica, la microbiómica y el ML, a lo que denominan las 3 M's, serán una herramienta clave para el diagnóstico y pronóstico temprano del COVID-19, así como para realizar predicciones de riesgo, estratificación, manejo de pacientes, y toma de decisiones, conduciendo a la medicina de precisión.

Fraser D., et al. encontraron niveles de metabolitos alterados, y fueron capaces de construir un modelo de ML con una exactitud del 98 % para el pronóstico y un 100 % de exactitud para la predicción de fallecimiento [9], y aunque en este estudio se utilizó un número limitado de pacientes (30 pacientes divididos en 3 grupos), muestran el potencial para la combinación de la metabolómica y el ML para el COVID-19, sobrepasando la necesidad de pruebas rt PCR [21].

3. Métodos y materiales

3.1. Conjunto de datos

El conjunto de datos utilizado en este estudio fue recolectado por López, Y., et al. y se compone de 37 personas confirmadas negativas que se sospechaban enfermas de COVID-19 debido a su contacto con individuos infectados, y 121 pacientes confirmados positivos. En este, se cuantificaron de forma precisa 110 metabolitos mediante la metabolómica dirigida, así como 13 citocinas/quimiocinas. El proceso de recolección de datos se describe detalladamente en [17], y el conjunto de datos completo se encuentra disponible bajo la licencia CC by 4.0 en [1].

3.2. Aprendizaje automático

El aprendizaje automático (ML, por sus siglas en inglés) es una rama de la inteligencia artificial que busca encontrar funciones desconocidas, relaciones, o estructuras entre un conjunto de variables de entradas y salidas a partir de un conjunto de datos, que muchas veces son difíciles de encontrar por algoritmos explícitos [24], mediante modelos que se ajustan al conjunto de datos.

[1] data.mendeley.com/datasets/x9tw3knwsd

El proceso de aprendizaje se realiza al encontrar los parámetros del modelo a partir del conjunto de datos de entrenamiento. Dependiendo de cómo se lleve este proceso de aprendizaje, un algoritmo de ML puede ser supervisado, no supervisado, semi-supervisado o aprendizaje profundo [5].

El ML es una de las tecnologías más prometedoras para la clasificación [12], y comúnmente se realiza mediante aprendizaje supervisado: el conjunto de datos consta tanto de las características de cada elemento a clasificar, como de la clase a la que pertenece.

En este caso, el algoritmo busca clasificar a cada elemento en alguna categoría y, a partir de un conjunto de D características $\vec{x} = (x_1, x_2, \cdots, x_D)$ [5]. En particular, para una clasificación binaria, la etiqueta y pertenece al conjunto $\{0, 1\}$. Aunque existe una gran variedad de algoritmos de ML para la clasificación binaria, dos de ellos utilizados comunmente son:

- **Árboles de decisión.** Un árbol de decisión, como es descrito por Kingsford, C. & Salzberg, S. L. [14], es un algoritmo que clasifica a un elemento mediante una serie de preguntas sobre sus características. Cada pregunta corresponde a un nodo, y cada posible respuesta apunta a un nodo hijo. Por lo tanto las preguntas forman un árbol. El elemento es etiquetado a una clase siguiendo la ruta de preguntas desde el primer nodo, la raíz, a un nodo sin hijos, una hoja, de acuerdo a las respuestas en cada nodo. La clase asignada al elemento es la asociada a la hoja que alcanza.
- **Bosques aleatorios.** Los bosques aleatorios (RF, por sus siglas en inglés), son otro algoritmo de ML, basado en los árboles de decisión. En este caso, se realiza un conjunto de árboles de decisión, y al final la clasificación se realiza mediante una votación (la clasificación es la clase más frecuente), o mediante algún método de promedio.

3.3. Selección de características

El problema de la seleccion de características (FS por sus siglas en inglés) puede ser establecido como la búsqueda de un subconjunto de d características de un total de D disponibles, que logran el mayor desempeño como características para realizar una clasificación [7], de acuerdo con alguna función de criterio.

Aunque se busca el conjunto de características óptimo, que maximice la función de criterio, actualmente no existe un método de FS capaz de resolver este problema en un tiempo razonable, especialmente cuando el conjunto de datos cuenta con varias decenas de características. Por lo tanto, se han desarrollado diferentes métodos que, aunque no garantizan encontrar la solución óptima, permiten obtener un conjunto subóptimo de características. Algunos métodos de FS utiilizados comunmente son:

- **Selección Secuencial hacia Adelante (SFS)**
  El método de Selección Secuencial hacia Adelante (SFS, por sus siglas en inglés) es un procedimiento que iterativamente trata de encontrar la nueva mejor característica que, junto a las demás, aumenta el desempeño. Iniciando con un conjunto con cero características, se agregan las características una a una: si el modelo mejora (o si se obtiene un resultado óptimo con la primera característica), esta se deja en el conjunto seleccionado, de lo contrario se elimina [7].

– **Algoritmos genéticos (GA)**
  El algoritmo de FS mediante Algoritmos Genéticos (GA, por sus siglas en ingles) es un método aleatorio guiado por una cierta medida de ajuste, e inspirado por el proceso natural de evolución. Cada posible conjunto de características (*individuo*) es representado por una cadena de D bits, $\alpha_1, \cdots, \alpha_D$, donde $\alpha_i = 1$ si la i-esima características se encuentra en el conjunto, y $\alpha_i = 0$ de lo contrario. En cada iteración del algoritmo (*generación*), un número fijo de posibles soluciones (*población*) es generado aplicando ciertos operadores genéticos (*recombinación*, *cruzamiento* y *mutación*) en un proceso estocástico. A medida que aumentan las generaciones, la población tiende a tener individuos con mejor desempeño, llegando a una solución subóptima.
– **Boruta**
  El algoritmo de boruta para la selección de características, implementado por primera vez para el lenguaje R en [15], está construido utilizando el algoritmo de bosques aleatorios. Este método consiste en agregar copias mezcladas aleatoriamente de todas las características (las características sombra) al conjunto de datos original, y utilizar un modelo de clasificación de bosques aleatorios con este nuevo conjunto de datos. Aplicando una métrica de importancia de características como la Exactitud de Disminución Media, se evalúa la importancia de cada característica. En cada iteración, el algoritmo de Boruta compara la importancia de cada característica con las características sombra, y elimina aquellas menos importantes, relativo a las características sombra [18].
– **LASSO**
  El algoritmo de Operador de Selección y Contracción Mínima Absoluta (LASSO, por sus siglas en inglés) es otro método de FS. Este modelo toma como base a la regresión lineal y agrega un término de regularización que penaliza la suma del valor absoluto de los coeficientes de regresión, L1, que forza a estos coeficientes predictorios a tender a cero. Para el proceso de FS, las variables que aún tienen un coeficiente diferente a cero después de aplicar la regularización son seleccionadas para el modelo [8].

3.4. Evaluación

Para la evaluación de los modelos existe una gran variedad de métricas. En particular, utilizamos la *exactitud* (probabilidad de que el algoritmo clasifique a alguna instancia correctamente), la *sensibilidad* (probabilidad de que el valor predicho sea correcto dado que el valor real es positivo), la *especificidad* (probabilidad de que el valor predicho sea correcto dado que el valor real es negativo) y la *precisión* (probabilidad de que la clasificación sea correcta, dado que el algoritmo hizo una clasificación como positivo) [5]. Estas métricas pueden ser obtenidos de la siguiente forma:

$$\text{Exactitud} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}, \tag{1}$$

$$\text{Sensibilidad} = \frac{T_P}{T_P + F_N}, \tag{2}$$

$$\text{Especificidad} = \frac{T_N}{T_N + F_P}, \tag{3}$$

$$\text{Precisión} = \frac{T_P}{T_P + F_P}, \tag{4}$$

donde T_P, T_N, F_P y F_N es el numero de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos, respectivamente. Sin embargo, si el número de instancias positivas y negativas es muy diferente, la exactitud no es una buena métrica. Para ello, consideramos la exactitud balanceada obtenida mediante la media aritmética de la sensibilidad y la especificidad. En el caso en el que el número de instancias positivas y negativas es igual, su valor es igual al de la exactitud normal. Además, utilizamos el f1 que se obtiene como la media armónica de la precisión y la sensibilidad:

$$\text{f1} = \frac{2}{\frac{1}{\text{Precisión}} + \frac{1}{\text{Sensibilidad}}}. \tag{5}$$

Con los cuatro valores anteriores (T_P, T_N, F_P y F_N) se obtiene la matriz de confusión que se forma al acomodar los valores en una matriz de 2×2 de la siguiente forma:

		Predicción	
Valor real		Verdaderos Positivos (T_P)	Falsos Negativos (F_N)
		Falsos Positivos (F_P)	Verdaderos Negativos (T_N)

Para tareas de clasificación binaria (0 o 1), cuando el resultado del modelo corresponde a la probabilidad de pertenecer a la clase positiva (1), se pueden obtener diferentes valores de sensibilidad y especificidad de acuerdo al umbral utilizado (la probabilidad mínima para que el modelo clasifique a la instancia como positiva).

La curva ROC (figura 3.4) se forma al graficar $1 - \text{especificidad}$ contra la sensibilidad al variar el umbral entre 0 y 1. El área bajo la curva ROC (AUC) representa el grado o la medida en la que el modelo es capaz de distinguir entre las dos clases: un valor de 1 indica un modelo sin error, un valor de 0 indica que el modelo realiza una clasificación opuesta (clasifica a los 0's como 1's y viceversa), mientras que un 0.5 indica que el modelo no tiene ninguna capacidad para separar las clases.

3.5. Validación

Comúnmente, el proceso de validación de un modelo consiste en dividir el conjunto de datos en 2 subconjuntos, uno de prueba y uno de validación.

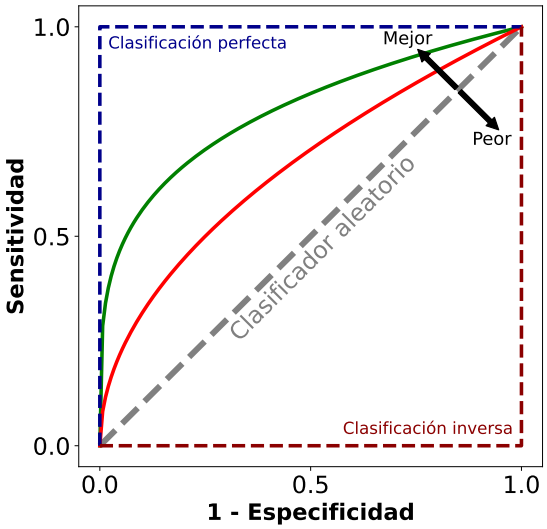


Fig. 1. Curva ROC. El área bajo la curva (AUC) representa el grado en el que el modelo es capaz de diferenciar entre dos clases. Un modelo perfecto (AUC=1) tendrá la curva azul, mientras que un modelo sin ninguna capacidad de clasificación (AUC=0.5) tendrá la curva gris. Una curva por debajo del clasificador aleatorio (curva gris, $\mathrm{AUC} < 0{,}5$), tenderá a hacer clasificaciones inversas (los 0's los clasifica como 1's y visceversa).

El primero de ellos es utilizado durante el entrenamiento, mientras que el segundo es utilizado durante la evaluación. Se le conoce como *generalización* a la capacidad de un modelo para realizar predicciones en datos (pacientes) que no se han utilizado durante el entrenamiento.

El método de validación cruzada dejando uno fuera (LOOCV, por sus siglas en inglés) consiste en realizar el proceso de validación N veces (donde N es el número total de instancias en el conjunto de datos), uno para cada una de las instancias. En cada ocasión, el modelo se prueba con una única instancia y se entrena con las $N - 1$ restantes, de manera que se utiliza la mayor cantidad de datos para el entrenamiento, y cada instancia pasa por la fase de validación.

El metodo de validación a ciegas consiste en separar un subconjunto de datos que no pasa por el proceso de generación y evaluación del modelo, sino que es utilizado para evaluar los modelos una vez ya generados. El conjunto de datos original se separa en dos subconjuntos: entrenamiento y prueba (train y test); el primero se utiliza para generar los modelos y evaluarlos mediante validación cruzada, y el segundo se utiliza una vez que los modelos ya están generados para estimar el rendimiento final.

3.6. Metodología seguida

El proceso general utilizado para la selección de los metabolitos y evaluación de los modelos de RF es el ilustrado en la figura 2, que es el correspondiente a una evaluación a ciegas.

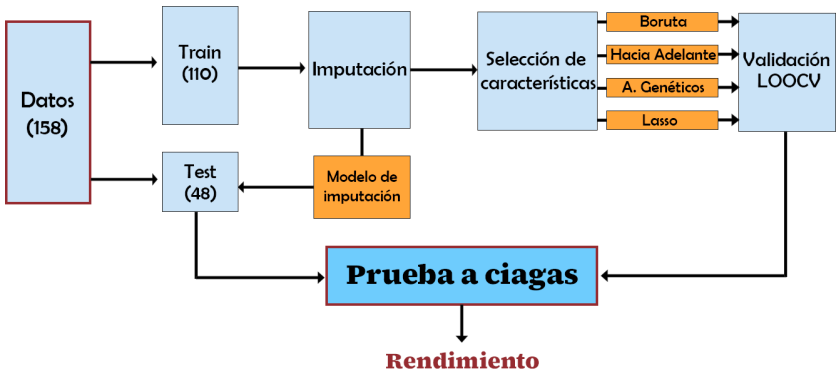


Fig. 2. Esquema general de la metodología seguida para realizar la selección de los metabolitos para el modelo de RF, y su evaluación final mediante una prueba a ciegas.

El conjunto de datos original, que consta de los datos de 158 pacientes, es dividido en 2 subconjuntos mediante una división de 70 % (110 pacientes para entrenamiento) - 30 % (48 pacientes para test).

El primer subconjunto de datos (train) es utilizado para generar el modelo de imputación de datos y para realizar los modelos de RF al seleccionar los metabolitos, los cuales son entrenados con este subconjunto y evaluados mediante una validación cruzada dejando uno fuera (LOOCV).

La validación del modelo a ciegas se realiza utilizando el segundo subconjunto de datos (test), al cual se le realiza la imputación de datos faltantes utilizando el modelo generado con el primer conjunto, y luego es utilizado para evaluar los modelos previamente entrenados con los datos de entrenamiento (train), con lo que se obtiene la estimación del rendimiento final de los modelos.

4. Resultados

Todos los resultados, mostrados en esta sección, fueron obtenidos mediante las cuatro técnicas de selección de características que se encuentran en la sección 3.3, utilizando de base un modelo de bosques aleatorios con 500 árboles de decisión, y el f1 como métrica de evaluación.

4.1. Selección del mejor conjunto de metabolitos

En conjunto, un total de 35 metabolitos diferentes fueron seleccionados por las 4 técnicas de selección de características utilizadas. La figura 3 muestra los metabolitos seleccionados con cada una de las técnicas. Entre ellos, destacan 3 que fueron seleccionados por 3 de las 4 tecnicas: el lysoPC a C26:0, el lysoPC a C14:00 y el radio quinurenina/triptófano, mientras que 10 fueron seleccionados por 2 de las 4 técnicas.

De los cuatro conjuntos seleccionados, el generado mediante selección hacia adelante es el que contiene la menor cantidad de características (5), seguido del generado mediante la técnica de LASSO (6).

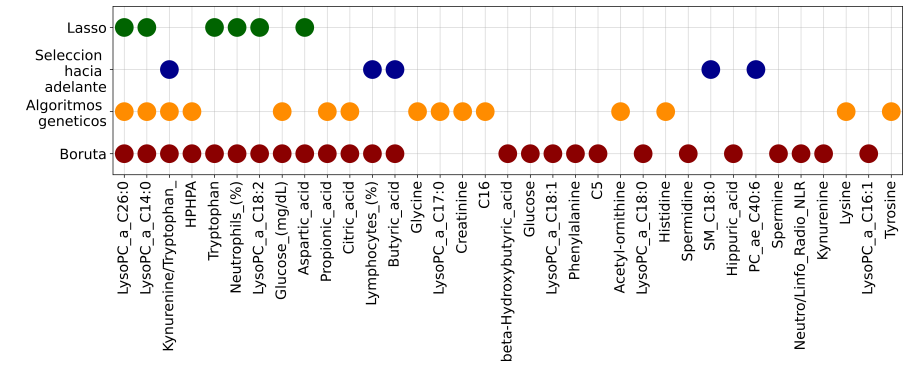
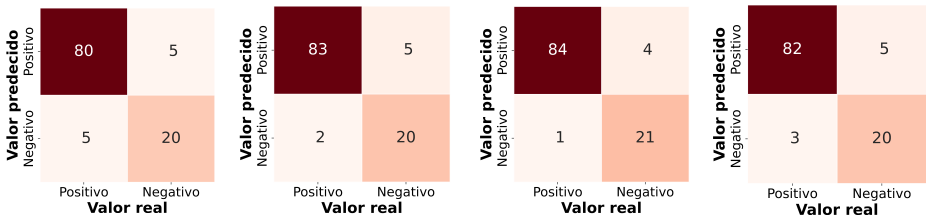


Fig. 3. Conjuntos de metabolitos seleccionados por las 4 técnicas de selección de características.



a) Boruta b) SFS c) GA d) LASSO

Fig. 4. Matrices de confusión para los 4 conjuntos de metabolitos seleccionados mediante técnicas de selección de características al realizar LOOCV en el conjunto de entrenamiento.

En cambio, las técnicas de Algoritmos Genéticos y Boruta generan los conjuntos más grandes, con 15 y 25, respectivamente, haciéndolos los menos prácticos para su uso como herramientas de diagnóstico, en cuestión de datos de laboratorio requeridos.

4.2. LOOCV en el conjunto de entrenamiento

La evaluación de los 4 conjuntos de metabolitos se realizó mediante LOOCV. La figura 4 muestra la matriz de confusión de los 4 modelos obtenida utilizando el conjunto de entrenamiento (con el cual se realizó la selección de características).

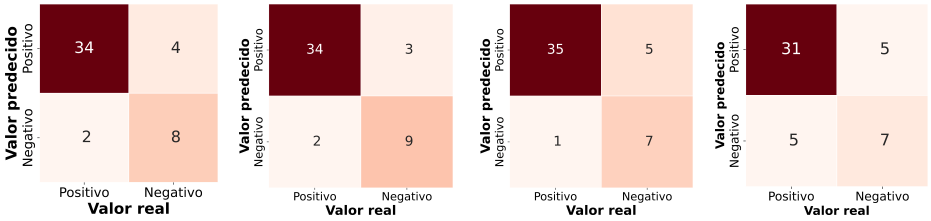
La tabla 1 muestra las diferentes métricas obtenidas a partir de los modelos anteriores. A pesar de la diferencia entre los conjuntos de metabolitos utilizados para cada modelo, los resultados de los 4 modelos son similares en términos de f1, siendo el obtenido por Algoritmos Genéticos el que alcanza un mejor resultado tanto en f1 como en exactitud balanceada.

4.3. Validación a ciegas

Para la validación a ciegas se utilizó el 30 % de los datos del conjunto original, y se utilizó el 70 % restante (el mismo conjunto con el que se realizó la selección de características) tanto para la imputación de datos faltantes como para el entrenamiento de los modelos.

Tabla 1. Métricas para la evaluación de los modelos generados con los 4 conjuntos de metabolitos obtenidos utilizando las técnicas de selección de características en el conjunto de entrenamiento por LOOCV.

	Lasso	**SFS**	**Boruta**	**GA**
Precisión	0.9647	0.9765	0.9412	0.9882
sensibilidad	0.9425	0.9432	0.9412	0.9545
f1	0.9534	0.9595	0.9411	0.9711
Exactitud balanceada	0.9272	0.9363	0.9091	0.9545
Especificidad	0.8696	0.9090	0.8000	0.9545
ROC_AUC	0.9572	0.9435	0.9689	0.9849



a) Boruta b) SFS c) GA d) LASSO

Fig. 5. Matrices de confusión para los 4 conjuntos de metabolitos seleccionados mediante técnicas de selección de características al realizar validación a ciegas.

La figura 5 muestra las matrices de confusión de los 4 conjuntos de metabolitos seleccionados, y en la tabla 2 se muestran las métricas de evaluación de los modelos. Además, en la figura 6 se muestran las curvas ROC tanto para el conjunto de entrenamiento como para la evaluación a ciegas.

Durante la validación a ciegas, el mejor modelo basado en términos de f1 y exactitud balanceada es el obtenido por selección hacia adelante, sin embargo, su rendimiento está muy cercano al obtenido mediante algoritmos genéticos (mejor rendimiento en el conjunto de entrenamiento) que se encuentra como el segundo mejor.

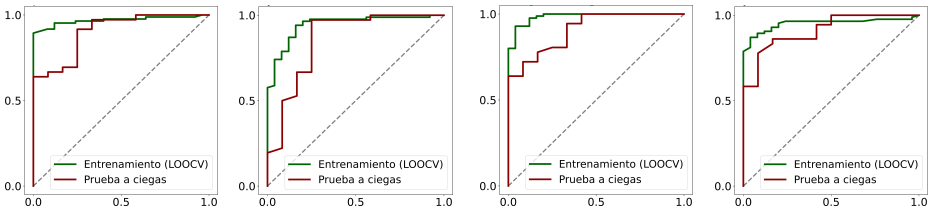
5. Conclusiones

Los resultados obtenidos en el presente trabajo de investigación muestran el desempeño del modelo de bosques aleatorios con niveles de metabolitos como características para el diagnóstico del COVID-19 en pacientes sospechosos, mostrando el alto potencial de la combinación de la metabolómica con el ML para el diagnóstico de esta enfermedad.

El conjunto con mejor desempeño en los datos de entrenamiento, en términos de la exactitud balanceada y el f1, fue el obtenido mediante algoritmos genéticos que consta de 15 metabolitos, y alcanza un f1 de 0.971 y una exactitud balanceada de 0.954. Al realizar la prueba ciega con los 48 pacientes, este conjunto obtiene el segundo mejor desempeño con un f1 de 0.921 y una exactitud de 0.875, muy cercano al obtenido por selección hacia adelante.

Tabla 2. Métricas para la evaluación de los modelos generados con los 4 conjuntos de metabolitos obtenidos utilizando las técnicas de selección de características al realizar validación a ciegas.

	Lasso	**SFS**	**Boruta**	**GA**
Precisión	0.8611	0.9444	0.9444	0.9722
sensibilidad	0.8611	0.9189	0.8947	0.8750
f1	0.8611	0.9315	0.8189	0.9211
Exactitud balanceada	0.7917	0.8958	0.8750	0.8750
Especificidad	0.5833	0.8182	0.8000	0.8750
ROC_AUC	0.9236	0.8576	0.9028	0.9100



a) Boruta b) SFS c) GA d) LASSO

Fig. 6. Curvas ROC para los 4 conjuntos de metabolitos, tanto en el conjunto de entrenamiento (obtenidos mediante LOOCV) como en la validación a ciegas.

En cambio, el conjunto obtenido por selección hacia adelante consta únicamente de 5 características y alcanza un f1 de 0.931 en la prueba a ciegas y una exactitud balanceada de 0.896, muy cercano al rendimiento obtenido al realizar LOOCV en el conjunto de entrenamiento (0.936 y 0.900, respectivamente).

A pesar del limitado número de pacientes utilizados para los modelos, el rendimiento obtenido en la prueba a ciegas es muy similar al obtenido con el conjunto de entrenamiento, lo cual puede observarse tanto al comparar las tablas 1 y 2 como al observar las curvas ROC, mostrando que los modelos tienen un alto nivel de generalización.

Dado el alto rendimiento mostrado en los resultados por los modelos, tanto en el conjunto de entrenamiento y especialmente al realizar la validación a ciegas, se propone el uso de estos conjuntos de metabolitos como biomarcadores de diagnóstico para el COVID-19, así como la futura implementación de diagnóstico automático utilizando tanto bosques aleatorios como otros modelos de ML, entrenados y evaluados con un conjunto de datos de más pacientes.

Agradecimientos. Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico brindado mediante la “Beca Nacional para Estudios de Posgrado” y por el desarrollo del proyecto “Paradigmas y Controversias de la Ciencia 2022” de número 319503 con el que fue obtenido el conjunto de datos, gracias a los cuales se está desarrollando el trabajo de investigación del cual surge el presente artículo.

Referencias

1. Agrebi, S., Larbi, A.: Use of artificial intelligence in infectious diseases. Artificial Intelligence in Precision Health, pp. 415–438 (2020) doi: 10.1016/B978-0-12-817133-2.00018-5
2. Ali, I., Alharbi, O. M.: Covid-19: Disease, management, treatment, and social impact. Science of The Total Environment, vol. 728, pp. 138861 (2020) doi: 10.1016/j.scitotenv.2020.138861
3. Asundi, A., O'Leary, C., Bhadelia, N.: Global COVID-19 vaccine inequity: The scope, the impact, and the challenges. Cell Host & Microbe, vol. 29, no. 7, pp. 1036–1039 (2021) doi: 10.1016/j.chom.2021.06.007
4. Azad, A. K., Hakim, A., Hasan-Sohag, M. M., Rahman, M.: Metabolomics in clinical diagnosis, prognosis, and treatment of infectious diseases. Metabolomics, pp. 71–119 (2023) doi: 10.1016/B978-0-323-99924-3.00003-0
5. Burkov, A.: The hundred-page machine learning book. vol. 1 (2019)
6. Davenport, T., Kalakota, R.: The potential for artificial intelligence in healthcare. Future healthcare journal, vol. 6, no. 2, pp. 94–98 (2019) doi: 10.7861/futurehosp.6-2-94
7. Ferri, F. J., Pudil, P., Hatef, M., Kittler, J.: Comparative study of techniques for large-scale feature selection. Machine Intelligence and Pattern Recognition, vol. 16, pp. 403–413 (1994) doi: 10.1016/B978-0-444-81892-8.50040-7
8. Fonti, V., Belitser, E.: Feature selection using lasso. VU Amsterdam Research Paper in Business Analytics, vol. 30, pp. 1–25 (2017)
9. Fraser, D. D., Slessarev, M., Martin, C. M., Daley, M., Patel, M. A., Miller, M. R., Patterson, E. K., O'Gorman, D. B., Gill, S. E., Wishart, D. S., Mandal, R., Cepinskas, G.: Metabolomics profiling of critically ill coronavirus disease 2019 patients: Identification of diagnostic and prognostic biomarkers. Critical Care Explorations, vol. 2, no. 10 (2020) doi: 10.1097/CCE.0000000000000272
10. Hasan, M. R., Suleiman, M., Perez-Lopez, A.: Metabolomics in the diagnosis and prognosis of COVID-19. Frontiers in Genetics, vol. 12, pp. 721556 (2021) doi: 10.3389/fgene.2021.721556
11. Hemdan, E. E. D., El-Shafai, W., Sayed, A.: 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, pp. 1–13 (2022) doi: 10.1007/s12652-022-03732-0
12. Hossain, B., Morooka, T., Okuno, M., Nii, M., Yoshiya, S., Kobashi, S.: Surgical outcome prediction in total knee arthroplasty using machine learning. Intelligent Automation & Soft Computing, vol. 25, no. 1, pp. 105–115 (2019)
13. Jamil, S., Mark, N., Carlos, G., Cruz, C. S. D., Gross, J. E., Pasnick, S.: Diagnosis and management of COVID-19 disease. American Journal of Respiratory and Critical Care Medicine, vol. 201, no. 10, pp. P19–P20 (2020) doi: 10.1164/rccm.2020C1
14. Kingsford, C., Salzberg, S. L.: What are decision trees? Nature biotechnology, vol. 26, no. 9, pp. 1011–1013 (2008) doi: 10.1038/nbt0908-1011
15. Kursa, M. B., Rudnicki, W. R.: Feature selection with the Boruta package. Journal of Statistical Software, vol. 36, no. 11, pp. 1–13 (2010) doi: 10.18637/jss.v036.i11
16. Lin, D. Y., Gu, Y., Wheeler, B., Young, H., Holloway, S., Sunny, S. K., Moore, Z., Zeng, D.: Effectiveness of COVID-19 vaccines over a 9-month period in North Carolina. The New England Journal of Medicine, vol. 386, no. 10, pp. 933–941 (2022) doi: 10.1056/NEJMoa2117128
17. López-Hernández, Y., Monárrez-Espino, J., Herrera-van Oostdam, A. S., Castañeda-Delgado, J. E., Zhang, L., Zheng, J., Oropeza-Valdez, J. J., Mandal, R.,

Ochoa-González, F. L., Borrego-Moreno, J. C., Trejo-Medinilla, F. M., López, J. A., Enciso-Moreno, J. A., Wishart, D. S.: Targeted metabolomics identifies high performing diagnostic and prognostic biomarkers for COVID-19. Scientific Reports, vol. 11, no. 1, pp. 14732 (2021) doi: 10.1038/s41598-021-94171-y

18. Perlato, A.: Feature selection using boruta algorithm (2020) www.andreaperlato.com/mlpost/feature-selection-using-boruta-algorithm/
19. Pourkarim, F., Pourtaghi-Anvarian, S., Rezaee, H.: Molnupiravir: A new candidate for COVID-19 treatment. Pharmacology Research & Perspectives, vol. 10, no. 1 (2022) doi: 10.1002/prp2.909
20. Shahin, O. R., Alshammari, H. H., Taloba, A. I., Abd El-Aziz, R. M.: Machine learning approach for autonomous detection and classification of COVID-19 virus. Computers and Electrical Engineering, vol. 101, pp. 108055 (2022) doi: 10.1016/j.compeleceng.2022.108055
21. Tounta, V., Liu, Y., Cheyne, A., Larrouy-Maumus, G.: Metabolomics in infectious diseases and drug discovery. Molecular Omics, vol. 17, no. 3, pp. 376–393 (2021) doi: 10.1039/D1MO00017A
22. Who coronavirus (COVID-19) dashboard, https://covid19.who.int/
23. Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., Hu, Y., Tao, Z. W., Tian, J. H., Pei, Y. Y., Yuan, M. L., Zhang, Y. L., Dai, F. H., Liu, Y., Wang, Q. M., Zheng, J. J., Xu, L., Holmes, E. C., Zhang, Y. Z.: A new coronavirus associated with human respiratory disease in China. Nature, vol. 579, no. 7798, pp. 265–269 (2020) doi: 10.1038/s41586-020-2008-3
24. Zhang, C., Yao, J., Hu, G., Schøtt, T.: Applying feature-weighted gradient decent K-Nearest neighbor to select promising projects for scientific funding. CMC Computers, Materials & Continua, vol. 64, no. 3, pp. 1741–1753 (2020) doi: 10.32604/cmc.2020.010306
25. Zoabi, Y., Deri-Rozov, S., Shomron, N.: Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj Digital Medicine, vol. 4, no. 1, pp. 3 (2021) doi: 10.1038/s41746-020-00372-6