Predicción precoz de la supervivencia en pacientes ingresados por COVID-19 con una aproximación integrada con variables ómicas

Oscar Martínez González

Resumen — La identificación precoz de pacientes con alto riesgo de mortalidad por COVID-19 podría mejorar su pronóstico y optimizar la gestión de recursos en futuras pandemias. En este estudio, se utilizan herramientas de Big Data para integrar y analizar 4408 variables clínicas, de laboratorio, microRNAs y metabolómicas en una cohorte de 95 pacientes. Mediante un proceso de depuración en cuatro etapas —que incluyó reducción de dimensionalidad con PCA, LASSO, ELASTICNET y MOFA— se seleccionaron 75 variables para desarrollar algoritmos de predicción de mortalidad, posteriormente validados en un grupo independiente. Los resultados demostraron que un conjunto reducido de biomarcadores moleculares, medidos dentro de las primeras 72 horas de hospitalización, permite predecir con precisión el riesgo de fallecimiento en pacientes con COVID-19. Este enfoque multiómico, facilitado por técnicas de Big Data, ofrece una herramienta prometedora para la toma de decisiones clínicas tempranas.

Palabras clave COVID-19, SARS-Cov2, Big data, dimensionalidad, mortalidad, ómica

Abstract— Early identification of patients at high risk of COVID-19 mortality could improve their prognosis and optimize resource management in future pandemics. In this study, Big Data tools are used to integrate and analyze 4408 clinical, laboratory, microRNAs and metabolomic variables in a cohort of 95 patients. Through a four-stage cleaning process-which included dimensionality reduction with PCA, LASSO, ELASTICNET and MOFA-75 variables were selected to develop mortality prediction algorithms, subsequently validated in an independent group. The results demonstrated that a reduced set of molecular biomarkers, measured within the first 72 hours of hospitalization, accurately predicts the risk of death in patients with COVID-19. This multi-omics approach, facilitated by Big Data techniques, offers a promising tool for early clinical decision making.

Index Terms— COVID-19, SARS-Cov2, Big data, dimensionality, mortality, omics

1 Introducción

Hasta mayo del 2025 se han recogido 778 millones de casos de COVID-19 en todo el mundo, 281 millones en Europa, de los cuales 13.873 casos corresponden a los últimos 28 días[1]. Resulta difícil saber la morbilidad y mortalidad exacta que se le puede atribuir a esta patología dada la amplia y variada distribución mundial. En el año 2024 la Organización Mundial de la Salud reportó que la pandemia ocurrida entre el 2019 y el 2021 acabó con casi una década de avances en la mejora de la esperanza de vida mundial, reduciéndola en 1,8 años, hasta los 71,4 años, así como en la esperanza de vida sana a nivel mundial en 1,5 años hasta los 61,9 años en 2021[2].

El cuadro clínico va desde pacientes asintomáticos hasta el desarrollo de insuficiencia respiratoria aguda que precisa de ingreso en una unidad de cuidados intensivos, con necesidad de soporte vital avanzado con ventilación mecánica invasiva, terapia renal sustitutiva, etc... llegando incluso al fallecimiento[3]. Dada la importante carga de trabajo y recursos que supone para los diferentes sistemas sanitarios, así como para la evolución del paciente, resulta especialmente relevante el poder predecir la evolución y pronóstico vital de los mismos desde su ingreso. Esto permitiría anticipar, adaptar y mejorar el sistema, así como los resultados en estos pacientes. Es por ello que se han realizado multitud de estudios desarrollando escalas pronósticas con gran diversidad de tipos de variables como pueden ser epidemiológicas, clínicas, de laboratorio hospitalario, proteinómicas, metabolómicas, transcriptómicas y genómicas. Generalmente se han analizado utilizando un único tipo de variables (clínica y de laboratorio, epidemiológica, genómica, proteínomica y transcriptómica)[4], [5], [6], [7], [8], [9], [10], [11].

Con los avances desarrollados en las últimas décadas en análisis bioquímico y análisis de datos se plantea la posibilidad de trabajar y estudiar una gran cantidad de datos y diversidad de variables integradas[4], [12], [13], [14], [15], [16]. Esta situación puede generar dificultades como pueden ser la elevada proporción de datos con respecto a los casos (“maldición de la dimensionalidad”), la dificultad para la integración de los diferentes tipos de variables y sus pesos a la hora de generar algoritmos predictivos, o la posibilidad de sobreajuste, haciendo que los algoritmos no puedan ser utilizados en la práctica clínica real[17].

El objetivo del estudio es predecir la mortalidad de manera precoz en pacientes ingresados en un hospital por

- *E-mail de contacto: intensivator@yahoo.es*
- *Trabajo tutorizado por: Remo Suppi*
- *Curso: 2024-25*

enfermedad Covid 19, que pueda ser exportable a la práctica clínica, utilizando nuevas técnicas bioquímicas y de aprendizaje supervisado. Para ello se van a utilizar variables de tipo clínico, de laboratorio hospitalario, transcriptómica y metabolómica, que nos permitan generar algoritmos predictivos supervisados e intentando recurrir al menor número de variables posibles.

2 Material y Métodos

2.1 Pacientes

Estudio prospectivo realizado en 95 pacientes de 18 años o más de edad, ingresados por Covid-19 entre Marzo y Agosto del año 2020 en tres hospitales públicos de la Comunidad de Madrid: Hospital Universitario del Tajo, Hospital Universitario Infanta Leonor y Hospital Universitario Príncipe de Asturias. El estudio fue aprobado por el Comité de Ética del Instituto Carlos III. Todos los pacientes incluidos firmaron un consentimiento informado para dicho estudio. Para poder incluir al paciente debía presentar una PCR positiva para SARS-Cov2 y clínica compatible que hubiera precisado ingreso hospitalario hasta 72 h antes de la inclusión. Se excluyeron pacientes con sospecha de otra posible patología como causa del cuadro clínico. El criterio de fallecimiento fue haber muerto en los 90 días posteriores a la fecha de ingreso hospitalario.

2.2 Variables

Se creó una única base de datos como suma de cuatro bases de datos previas: clínica y demográfica; laboratorio hospitalario; análisis transcriptómico; y análisis metabolómico. Los estudios transcriptómicos y metabolómicos se realizaron por el Centro Nacional de Microbiología del Instituto de Salud Carlos III.

2.3 Procedimiento experimental

Dependiendo del tipo de variable el procedimiento de adquisición fue diferente.

a) Variables clínicas y de laboratorio hospitalario: se recopilaron exhaustivamente de las historias clínicas mediante un formulario electrónico de recogida de datos creado con las herramientas de captura de datos electrónica REDCap. Las muestras de plasma se recogieron al ingreso hospitalario o en los primeros días tras la hospitalización y antes del tratamiento con terapias específicas para la COVID-19.

b) Análisis de miRNomas: se llevó a cabo mediante secuenciación de alto rendimiento. Para ello se aisló el ARN total, incluyendo los ARN pequeños, de 400 µl de plasma con el kit miRNeasy Serum Plasma Advanced (Qiagen, Hilden, Alemania). La calidad y cantidad del ARN se evaluaron con el Bioanalyzer 2100 y el kit Agilent RNA 6000 Nano. Se construyeron bibliotecas de ARN pequeños utilizando el programa NEBNext Multiplex Small RNA Library Prep para Illumina (New England Biolabs, Ipswich, MA, EE. UU.) en el Parque Científico de Madrid (España). A continuación, se secuenciaron los ARN pequeños en NovaSeq 6000, Illumina, en el Centro Nacional de Microbiología (Majadahonda, España), con un rendimiento estimado de más de 50 millones de lecturas por muestra. Se realizó una comprobación de calidad de las lecturas con FastQC (v.0.11.3) y se recortaron las secuencias adaptadoras con cutadapt (v.1.13). Posteriormente, las lecturas se procesaron con miRDeep2 (v.0.0.7) para identificar y cuantificar los miRNA conocidos del genoma humano de referencia (GRCh38) y miRBase (v2.0).

c) Análisis metabolómico no dirigido: las muestras de plasma fueron inactivadas con una mezcla (-20 °C) de MeOH:EtOH (1:1, v/v) frío, agitadas en vórtex durante 1 minuto, incubadas en hielo durante 5 minutos y posteriormente centrifugadas durante 20 minutos a 16,000 xg a 4 °C. El sobrenadante resultante se almacenó a -80 °C hasta su análisis. Debido a las amplias diferencias en las propiedades físico-químicas de los metabolitos, se utilizaron tres plataformas analíticas para aumentar la cobertura metabolómica: cromatografía de gases acopada a espectrometría de masas (GC-MS) centrada en moléculas pequeñas que pueden volverse volátiles mediante derivatización, electroforesis capilar acoplada a espectrometría de masas (CE-MS) enfocada en compuestos polares e iónicos y cromatografía de líquidos acoplada a espectrometría de masas (LC-MS) orientada a metabolitos semipolares y lipofílicos. Para evaluar la calidad de los datos de cada plataforma, se procesaron muestras de control de calidad (QC), utilizando mezclas de volúmenes iguales de cada muestra correspondiente. Además, se analizaron dos soluciones en blanco, una al inicio y otra al final de cada secuencia analítica.

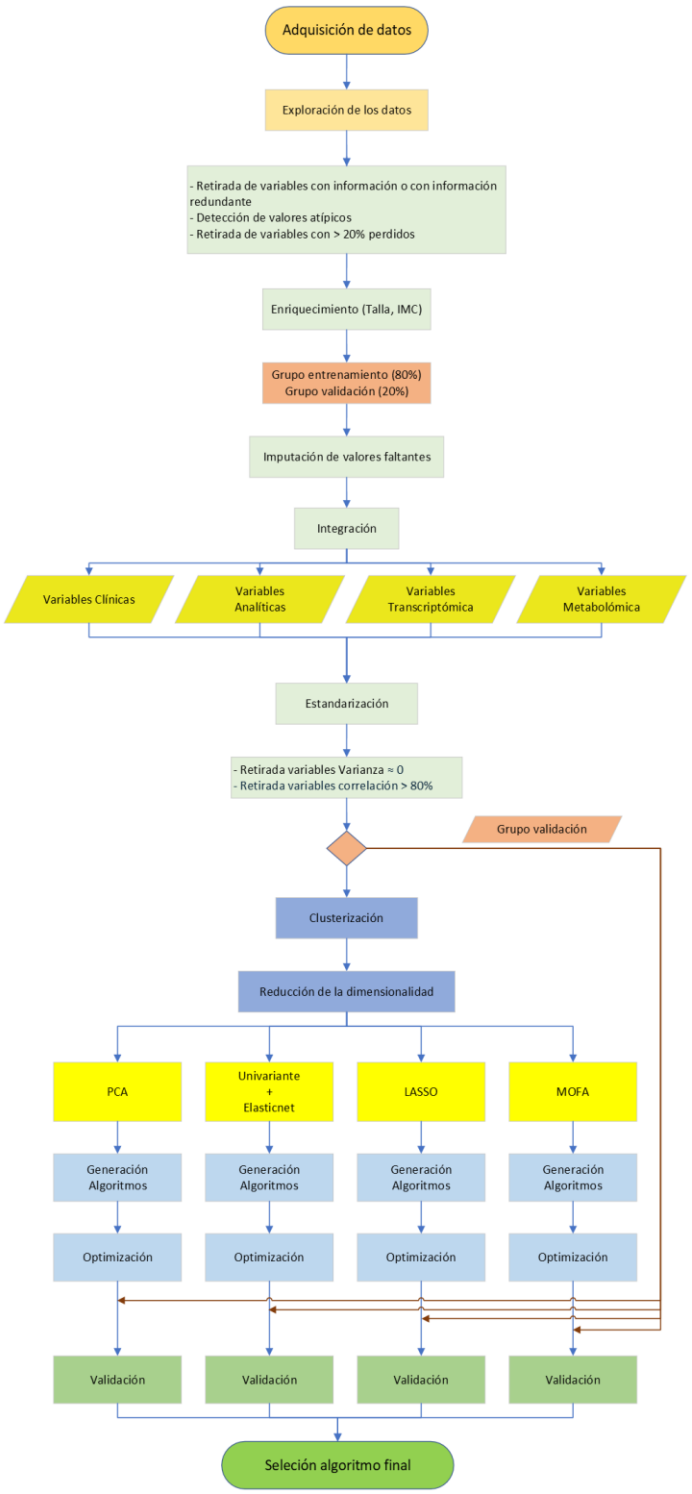
Para los análisis, se descartaron aquellos metabolitos con baja reproducibilidad (coeficiente de variación [CV] en los controles de calidad superior al 30%) y aquellas señales no presentes en al menos el 70 % de las muestras de algún grupo. En el caso de GC-MS, la concentración final de cada metabolito se normalizó también según la abundancia del estándar interno (IS). Finalmente, se corrigió la caída de intensidad en los controles de calidad, y las matrices resultantes se utilizaron para el análisis estadístico. Las características metabólicas obtenidas mediante GC-MS y CE-MS se encuentran identificadas. En cuanto a los datos correspondientes a LC-MS, tras la realización del análisis estadístico, se llevará a cabo la identificación de los compuestos se (pendiente en el momento de redacción de este trabajo).

2.4 Análisis estadístico de los datos

El análisis de datos se realizó en Python (3.10.16) utilizando las bibliotecas Pycaret (3.3.2), Scikit-learn (1.4.2), Pandas (2.1.4), Numpy (1.26.4) y Scipy (1.11.4).

Análisis descriptivo y exploración inicial. Se ha realizado un análisis descriptivo de la población de estudio, de los pacientes vivos y de los pacientes fallecidos. Para la exploración estadística inicial de los datos se ha realizado el test de Mann Whitney o t de Student para las variables continuas, y el test de Chi Cuadrado o test de Fischer para las variables categóricas. Posteriormente se realizaron técnicas de aprendizaje no supervisado con K-means y DBSCAN como análisis exploratorio inicial. Dado que no era posible generar un número de grupos razonable se utilizaron Uniform Manifold Approximation and Projection (UMAP) y t-distributed Stochastic Neighbor Embedding (t-SNE). Estos son algoritmos de reducción de dimensionalidad no lineal. Su objetivo principal es tomar datos de alta dimensionalidad (muchas características) y proyectarlos en un espacio de baja dimensionalidad (típicamente 2D o 3D) de manera que se preserve la estructura local o global de los datos originales. Son especialmente útiles para la visualización de datos complejos, permitiendo identificar agrupaciones (clusters), patrones o la forma general de la distribución de los datos.

Figura 1. Flujo de trabajo de los datos



Preprocesado de los datos. Se generó un proceso de trabajo con los datos con pasos iniciales comunes y posteriormente separados (Figura 1). Antes de comenzar con el análisis de los datos se procedió a valorar los valores límites y retirar variables en base a dicha información, las que presentaban valores perdidos > 20%, varianza cercana a 0, ó no aportaban información o eran redundantes. Se procedió a la generación de un grupo de entrenamiento (80%) y otro de validación (20%), estratificado para la variable "fallecimiento". Posteriormente se imputaron los valores por la media para variables numéricas y proximidad para categóricas, y se estandarizó por grupos.

Reducción de la dimensionalidad. Tras la integración de las variables restantes se retiraron las variables con varianza próxima a 0 ó correlación > 80%. En el grupo de entrenamiento se procedió a utilizar 4 técnicas de reducción de la dimensionalidad diferentes. Principal Component Analysis (PCA), regresión con regularización Elasticnet, regresión con regularización LASSO y Multi-Omics Factor Analysis (MOFA). Este último consiste en un modelo de análisis factorial que proporciona un marco general para la integración de conjuntos de datos multiómicos de forma no supervisada. Intuitivamente, MOFA puede considerarse una generalización versátil y estadísticamente rigurosa del análisis de componentes principales a datos multiómicos. Dadas varias matrices de datos con mediciones de múltiples tipos de datos ómicos en el mismo conjunto de muestras o en conjuntos superpuestos, MOFA infiere una representación interpretable de baja dimensión en términos de unos pocos factores latentes. Estos factores aprendidos representan las fuentes impulsoras de la variación entre las modalidades de datos, lo que facilita la identificación de estados celulares o subgrupos de

enfermedades.

Una vez aplicadas las diferentes técnicas de reducción de la dimensionalidad se escogieron las 75 variables con más peso en cada uno de los grupos de entrenamiento depurados y reducidos. Dado que el grupo de entrenamiento constaba de 76 casos se decidió que las variables a utilizar para generar los algoritmos fueran como máximo el número de pacientes -1.

Generación de algoritmos. Para la generación de los algoritmos de aprendizaje automático supervisado se utilizó exclusivamente el grupo de entrenamiento. Utilizando la librería Pycaret (pycaret.org) de Automated Machine Learning se crearon y optimizaron con validación cruzada de manera automática dentro del mismo grupo de entrenamiento y posteriormente se evaluaron los resultados en el grupo test. Esta librería automatiza los flujos de trabajo de aprendizaje automático y permite modelar multitud de variables de manera sencilla, creando algoritmos y sugerencias para la optimización de los mismos a través de diferentes técnicas (árboles, importancia de la permutación, regresión logística, k-Nearest Neighbors), lo que ayuda y facilita la optimización. Se utilizó esta herramienta como generador de algoritmos y optimización de los mismos con el *Area Under the Curve* (AUC), seguida del F1 score.

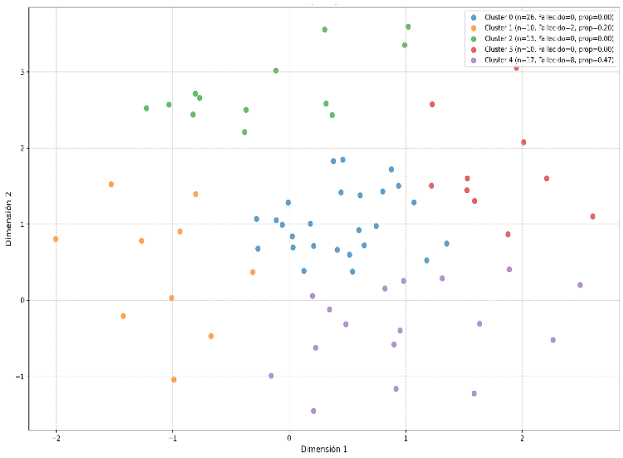
Selección de algoritmos. Para seleccionar los algoritmos se valoró la relación entre el número de variables y el ajuste de los mismos en el grupo de validación, de manera que facilite la traslación a la práctica clínica. Para ello se consideró el parámetro número de eventos/número de variables.

3 Resultados

Se incluyeron 95 pacientes con una mediana de edad de 60 años y un 44,2% de mujeres. Fallecieron 13 pacientes, 4 de ellos mujeres. Las principales características clínicas y de laboratorio se encuentran descritas en la tabla 1. De estas solo presentaban diferencias significativas en el análisis univariante entre los grupos vivos y fallecidos a los 90 días el porcentaje de neutrófilos, el porcentaje de linfocitos, la creatinina, la enzima lactato deshidrogenasa (LDH) y la proteína C reactiva (PCR).

Inicialmente se generó un análisis no supervisado de las diferentes vistas de datos (variables clínicas, variables de laboratorio, microRNA y metabolitos) para valorar la distribución de los pacientes fallecidos en los grupos generados de forma automática. Los métodos K-means y DBSCAN no fueron capaces de generar grupos ante la gran dimensionalidad, por lo que se aplicaron técnicas de reducción de la dimensionalidad para entornos no supervisados (UMAP y t-SNE), generándose grupos con grandes diferencias en el número y proporción de pacientes fallecidos (Figura 2).

Figura 2. Clústeres generados con K-means sobre t-SNE con todas las variables



Tras la exploración inicial que confirma la existencia de diferencias en los datos entre pacientes vivos y fallecidos a los 90 días, se estimaron el número de componentes principales necesarios para explicar el porcentaje de la varianza con los diferentes componentes y en las diferentes vistas principales o con el número de componentes principales necesarios optimizados por regresión lineal, por SVM, por ADABOOST y finalmente por Random Forest (tabla 2). El número de componentes principales para las diferentes vistas varió entre 16, en el grupo de variables de laboratorio, y 65. En la figura 3 se muestran los AUC con 2 componentes principales con las diferentes vistas. Curiosamente el modelo con 2 componentes principales con las variables de laboratorio fue el que mejor AUC presentó, siendo de 0,85 y mejor que el resto de modelos con todas las variables.

Posteriormente se realizaron las diferentes técnicas de reducción de la dimensionalidad (LASSO, ELASTICNET y MOFA) y tras seleccionar las 75 variables con más peso en cada una de ellas se procedió a generar los modelos (regresión logística, support vector machine, Linear Discriminant Analysis, Ada Boost Classifier y Random forest). En la tabla 3 se muestran las diferentes técnicas de reducción de la dimensionalidad con las métricas y características principales de los diferentes modelos. En la figura 3 se muestran las AUC. El modelo con mejor AUC fue el resultante de aplicar la técnica LASSO y generar el modelo con random forest con 67 variables, con un AUC de 1,00. Los que precisaron menos variables fueron la regresión logística y el SVM tras aplicar MOFA, precisando únicamente 2 variables.

Tabla 1. Características generales

	Total	Vivos	Fallecidos	p
	n = 95	n = 82	n = 13	
Edad	60 [40-73]	60 [40 - 73]	63 [59 - 76]	0.452
Mujer (%)	42 (44.2)	38 (46.34)	4 (30.77)	0.431
Clínica				
Frecuencia respiratoria (rpm)	21 [18 - 30]	20 [18 - 28]	25 [20 - 32]	0.053
Presión arterial sistólica (mmHg)	100 [91 - 111]	100 [92 - 110]	94 [82 - 115]	0.183
Presión arterial diastólica (mmHg)	59 [51 - 69]	60 [52 - 68]	53 [43 - 71]	0.166
Saturación O2 (%)	94 [92 - 97]	94 [92 - 97]	93 [92 - 94]	0.151
Temperatura	38.0 [37.3 -38.6]	38.0 [37.3 - 38.5]	38.0 [37.4 - 38.8]	0.581
Diarrea (%)	33 (34.74)	31 (38.27)	2 (15.38)	0.179
Naúseas (%)	18 (18.95)	18 (22.22)	0 (0)	0.114
Dolor abdominal (%)	15 (15.79)	12 (14.81)	3 (23.08)	0.702
Artralgias/mialgias (%)	26 (27.37)	24 (29.63)	2 (15.38)	0.566
Disnea (%)	61 (64.21)	50 (61.73)	11 (84.62)	0.196
Anosmia /Ageusia (%)	7 (7.37)	6 (7.41)	1 (7.69)	0.84
Rinitis (%)	4 (4.21)	4 (4.94)	0 (0)	0.267
Tos (%)	63 (66.32)	58 (70.73)	5 (38.46)	0.068
Antecedentes				
Embarazada (%)	2 (2.11)	2 (2.47)	0 (0)	1
Fumador (%)	8 (8.42)	8 (9.88)	0 (0)	0.468
Hipertensión arterial (%)	41 (43.16)	32 (39.51)	9 (69.23)	0.088
Diabetes (%)	19 (0.20)	18 (22.22)	1 (7.69)	0.401
Insuficiencia renal crónica (%)	12 (12.63)	8 (9.88)	4 (30.77)	0.1
Tumor oncohematológico (%)	3 (3.16)	2 (2.47)	1 (7.69)	0.525
Neoplasia solida (%)	5 (5.26)	4 (4.94)	1 (7.69)	0.293
Cardopatía (%)	17 (17.89)	13 (16.05)	4 (30.77)	0.375
Enfermedad pulmonar (%)	15 (15.79)	11 (13.58)	4 (30.77)	0.276
Enfermedad neurológica (%)	14 (14.74)	10 (12.35)	4 (30.77)	0.211
Enfermedad autoinmune (%)	4 (4.21)	4 (4.94)	0 (0)	0.715
Tratamientos				
VIH positivo (%)	2 (2.11)	2 (2.47)	0 (0)	1
Corticoides (%)	47 (49.47)	38 (46.91)	9 (69.23)	0.232
Anticoagulación (%)	50 (52.63)	43 (53.09)	7 (53.85)	1
Tozilizumab (%)	23 (24.21)	19 (23.46)	4 (30.77)	0.824
AINES (%)	4 (4.21)	2 (2.47)	2 (15.38)	0.05
Analítica				
Neutrofilos (% leucocitos)	76 [64 - 86]	75 [63 -82]	85 [80 - 87]	0.013
Linfocitos (% leucocitos)	17 [10 - 26]	19 [12 - 27]	9 [7 - 10]	0.002
Plaquetas	244.5 [174.5 - 318.5]	242.0 [175.5 - 315.0]	262.0 [157.0 - 365.0]	0.862
Dímero D	950 [500 - 2460]	780 [485 - 1480]	3500 [2980 - 5400]	0.193
Creatinina (mg/dl)	0.90 [0.74 - 1.14]	0.89 [0.73 - 1.04]	1.24 [0.89 - 2.41]	0.003
LDH	248 [206 - 362]	242 [192 - 333]	384 [262 - 464]	0.004
GPT	32 [24 - 71]	35 [22 - 71]	29 [27 - 71]	0.614
PCR	52 [17 - 118]	36 [11 - 93]	131 [81 - 175]	0.001

Las variables que más peso presentaban en los diferentes algoritmos fueron principalmente metabolitos y alguno microRNAs. Destacan por su presencia en múltiples modelos con diferentes técnicas de reducción de la dimensionalidad los metabolitos '787.6278@12.23', '928.5713@7.13' y '823.6276@12'. El peso de las variables clínicas, las utilizadas a día de hoy en la práctica real, fue escaso. La raza, la presencia de tos como síntoma y el haber precisado tratamiento antibiótico antes de entrar en el estudio son las que presentan mayor importancia en los diferentes algoritmos.

Para decidir que modelos se consideraron adecuados para su aplicación en la práctica clínica se excluyeron todos aquellos que precisaran más de 15 variables o tuvieran un AUC mayor de 0,95 por el elevado riesgo de sobreajuste. De un total de 23 posibles modelos, incluyendo los generados con PCA, quedaron 11. Teniendo en cuenta la relación entre el número de variables y los resultados de las métricas en el grupo de validación se decidió que los modelos que presentaban una mejor relación entre el número de variables y las métricas resultantes para poder utilizarlos en la práctica clínica fueron el random forest con 2 variables (AUC 0,92) que es exactamente igual tras ELASTICNET y MOFA, la regresión logística con 4 variables tras reducir con ELASTICNET (AUC 0,90), y la regresión logística tras aplicar MOFA con 2 variables (AUC 0,89).

4 Discusión

A día de hoy todavía se sufren las consecuencias de la pandemia padecida en el año 2020. La variación de la esperanza de vida años más tarde, la sobrecarga de los sistemas sanitarios si ocurriera nuevamente, la necesidad de predecir y elegir correctamente a que pacientes tratar y que recursos utilizar son cuestiones en las que se ha avanzado. A pesar de lo cual todavía quedan muchas preguntas. Si bien existen multitud de estudios prediciendo la mortalidad en el paciente con infección por Covid-19, el presente estudio es el que utiliza mayor variedad y número de variables. Ello supone una buena representación de lo que implica la integración de nuevas técnicas diagnósticas junto con otras clásicas aplicando nuevas herramientas en inteligencia artificial y big data. Por todos estos motivos esta muestra resulta adecuada para abordar los diferentes problemas que surgen de esta integración.

La decisión de generar un algoritmo que se pudiera llevar a la práctica clínica hizo que se optara por utilizar modelos de aprendizaje supervisado y no de aprendizaje profundo, ya que con ello se simplifica la comprensión de los resultados e interpretación, facilitando su aplicación. También se da mucha importancia al número de variables, intentando optimizar la relación nº de eventos/nº de variables, de manera que se evite el sobreajuste y simplifique su uso. Cuanto mayor es el valor de este cociente mayor es el rendimiento del modelo en una muestra nueva tomada en la misma población, lo que implica disminuir el sobreajuste[18]. Este parámetro puede ser utilizado incluso para sustituir la necesidad de validación externa cuando es mayor de 20. En el caso particular del modelo elegido el valor fue de 4,5, insuficiente para evitar la evaluación en el grupo de validación, pero suficientemente elevado en relación al número de casos. Con estos condicionantes el modelo elegido como el que se ajustaba mejor para su aplicación fue

Figura 3. Curvas AUCROC de los diferentes modelos tras aplicar diferentes técnicas de reducción de dimensionalidad

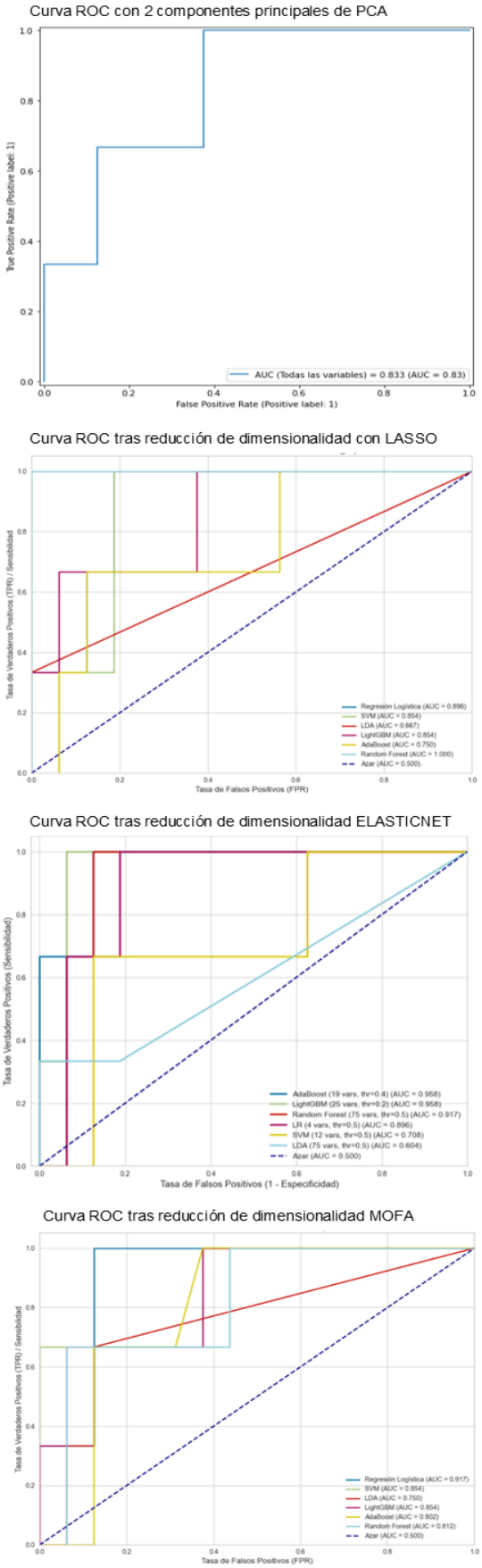


Tabla 2. Resultados con diferentes modelos de PCA

Modelo	Nº Componentes	Vista	AUC	F1		Recall		Precission	
				Balanceada	Fallecidos	Balanceada	Fallecidos	Balanceada	Fallecidos
2 Componentes Principales	2	Laboratorio	0,854	0,740	0,000	0,790	0,000	0,700	0,000
Optimizada Regresión logística	10	Laboratorio	0,833	0,740	0,000	0,790	0,000	0,700	0,000
Optimizada SVM	16	Metabolitos	0,813	0,770	0,000	0,840	0,000	0,710	0,000
Optimizada ADABOOST	54	microRNA	0,840	0,740	0,000	0,790	0,000	0,700	0,000
Otimizada Random forest	61	Metabolitos	0,830	0,740	0,000	0,790	0,000	0,700	0,000

Tabla 3. Resultados de los algoritmos tras los diferentes tipos de reducción de dimensionalidad

Modelo	Nº variables	Accuracy	AUC	F1		Recall		Precission		Variables más importantes	Umbral
				Balanceada	Fallecidos	Balanceada	Fallecidos	Balanceada	Fallecidos		
LASSO											
Regresión logística	57	0,79	**0,90**	0,87	0,50	0,89	0,33	0,91	1,00	178.0335@8.75; 283.2642@6.09; hsa-miR-6781-5	0,50
SVM	56	0,89	0,85	0,87	0,50	0,89	0,67	0,91	0,40	raza; 592.9061@0.8243; 1009.6799@12.01	0,50
Linear Discriminant Analysis	57	0,89	0,67	0,87	0,50	0,89	0,33	0,91	1,00	tos; Antibiótico; 643.3687@0;97	0,50
Light Gradient Boosting Machine	42	0,84	0,85	0,83	0,40	0,84	0,33	0,82	0,50	hsa-miR-4510; 431.2201@1.12; 178.0335@8.75	0,30
Ada Boost Classfier	3	0,84	0,75	0,85	0,57	0,84	0,67	0,86	0,50	788.5982@7.44; hsa-miR-4510; 431.2201@1.12	0,30
Random forest	67	0,88	**1,00**	0,90	0,75	0,89	1,00	0,94	1,00	431.2201@1.12; 788.5982@7.44; 1009.6799@12;01	0,40
Ridge Classifier	5	0,89	0,67	0,87	0,50	0,89	1,00	0,91	1,00	hsa-miR-92b-5p; hsa-miR-6781-5p; 178.0335@8.75	0,50
ELASTICNET											
Regresión logística	4	0,89	**0,90**	0,89	0,67	0,89	0,67	0,89	0,67	787.6278@12.23; 928.5713@7.13; 381.3363@4.4	0,50
SVM	12	0,84	0,71	0,77	0,00	0,84	0,00	0,71	0,00	774.654@11.8049; 690.685@12.2526; 368.2793@1.1874	0,50
Linear Discriminant Analysis	57	0,79	0,60	0,79	0,33	0,79	0,33	0,79	0,33	fallo renal; 537.4882@3.76; 683.5931@12.2	0,50
Light Gradient Boosting Machine	25	0,89	**0,96**	0,89	0,67	0,89	0,67	0,89	0,67	787.6278@12.23; 928.5713@7.13; 877.7846@11.91	0,20
Ada Boost Classfier	19	0,89	**0,96**	0,95	0,86	0,95	1,00	0,96	0,75	787.6278@12.23; 928.5713@7.13; hsa-let-7f-1-3p	0,40
Random forest	2	0,95	**0,92**	0,94	0,80	0,95	0,67	0,95	1,00	787.6278@12.23; 928.5713@7.13	0,50
Ridge Classifier	21	0,79	0,64	0,79	0,33	0,79	0,33	0,79	0,33	hsa-miR-1255a; 774.654@11.8049; 582.2808@2.25	0,50
MOFA											
Regresión logística	2	0,92	0,89	0,90	0,75	0,89	1,00	0,94	0,60	787.6278@12.23; hsa-miR-22-3p	0,50
SVM	2	0,85	0,84	0,87	0,50	0,89	0,33	0,91	1,00	hsa-miR-22-3p; 445.2691@0.93	0,80
Linear Discriminant Analysis	74	0,75	0,89	0,94	0,80	0,95	0,67	0,95	1,00	823.6276@12; 951.579@7.13; 928.5713@7.13	0,60
Light Gradient Boosting Machine	39	0,85	**0,95**	0,94	0,80	0,95	0,67	0,95	1,00	823.6276@12; 624.5565@11.27; 787.6278@12.23	0,50
Ada Boost Classfier	9	0,80	**0,95**	0,94	0,80	0,95	0,67	0,95	1,00	928.5713@7.13; 787.6278@12.23; 1026.6057@11.54	0,60
Random forest	2	0,95	**0,92**	0,94	0,80	0,95	0,67	0,95	1,00	787.6278@12.23; 928.5713@7.13	0,50
Ridge Classifier	4	0,67	0,89	0,87	0,50	0,89	0,33	0,91	1,00	787.6278@12.23; hsa-miR-22-3p; IP.10_FI	0,50

la regresión logística tras MOFA con 2 variables (AUC 0,89). Además, presentaba un valor de F1 score de 0,90 y una sensibilidad de 1 en el grupo de fallecidos, lo cual implica la ausencia de falsos negativos. Este es el motivo de su elección, ya que el evento es el fallecimiento y en ese caso se asume un menor ajuste en la clasificación en general a cambio de no dejar pacientes que van a fallecer fuera del grupo de riesgo. Llama la atención y es destacable que, a pesar de utilizar tres técnicas de reducción de la dimensionalidad diferentes, en dos de ellas el resultado en el algoritmo generado con random forest es exactamente igual y con unos resultados muy buenos. Esta congruencia refuerza la conclusión extraída.

Cuando se comparan las métricas de los resultados con la validación externa de diferentes modelos en poblaciones variadas estos se encuentran entre los que mejor ajuste presentan y con menor necesidad de variables para la realización de la predicción[19], [20]. Si bien para confirmar esta posibilidad se precisaría una validación externa en la población general.

El diseño de un estudio prospectivo para predecir la mortalidad hace que los datos recogidos sean consistentes para generar los algoritmos y medir el resultado objetivo. Este diseño asegura una proporción pequeña de valores perdidos, lo que evita las deficiencias relacionadas con el tratamiento de los mismos. Uno de los problemas de este tipo de estudios es el seguimiento de los pacientes, pero en este caso se pudo seguir a todos los pacientes hasta el día 90 tras el ingreso, ya que a fecha de creación de este artículo todavía se mantiene el seguimiento a todos los pacientes vivos[21].

Se decidió generar un grupo de entrenamiento y uno validación desde el inicio con el objetivo de aislar lo máximo posible un grupo del otro, suponiendo que con estas medidas las métricas obtenidas en el grupo de validación serían extrapolables al resto de la población representada, en nuestro caso pacientes infectados e ingresados en el hospital por COVID-19 en la Comunidad de Madrid. Los modelos se generaron con el grupo de entrenamiento y un subgrupo del mismo usado como grupo test. Esta medida puede suponer una pérdida de ajuste, pero probablemente habrá mejorado la aplicabilidad posterior. Que el estudio sea multicéntrico es otro de los puntos a favor de que los algoritmos generados aplicados en la misma población den resultados similares a los obtenidos en el grupo de validación.

El estudio presenta múltiples debilidades. La muestra es pequeña, lo que dificulta obtener mejores resultados sin riesgo de sobreajuste y una posible falta de representación de posibles grupos de riesgo. En general, los trabajos con múltiples ómicas suelen realizarse en muestras pequeñas de pacientes debido a lo complejo y costoso de obtener estos datos. Hasta donde se sabe, este estudio es el que integra mayor número de variables por paciente y como ya se ha descrito tiene un elevado riesgo de presentar la "maldición de la dimensionalidad". Por otro lado, el evento a estudio es la mortalidad, que se encuentra desbalanceado en la muestra con respecto a la supervivencia (13,68% vs 86,32%), pero este desbalance representa la realidad[22]. Para ello se puede plantear en el futuro utilizar técnicas y/o herramientas para balancear el evento o generar pacientes sintéticos que lo presenten. Los pacientes pertenecen a un único sistema sanitario lo que dificulta la posibilidad de exportar los resultados a pacientes de otros sistemas diferentes. Finalmente, los pacientes de este estudio no se encontraban vacunados en el momento de ingresar, por lo que a la hora de generalizar estos datos en la actualidad hay que ser cautelosos y sería recomendable confirmar estos hallazgos en un grupo que incluyera pacientes vacunados. Estas últimas objeciones podrían comprobarse con la validación de un nuevo grupo amplio de pacientes representativos de los diferentes sistemas en la actualidad.

7 Conclusión

Los resultados demuestran que un reducido número de variables de tipo molecular obtenidas en las primeras 72 h tras el ingreso pueden predecir el fallecimiento en los 90 días posteriores al ingreso en pacientes hospitalizados por COVID-19 con mucha precisión. Con estos resultados se podría plantear la creación de sencillos paneles moleculares que nos sirvan para estratificar el riesgo de fallecimiento, mejorando el seguimiento y pronóstico de los pacientes, así como facilitando la optimización de los recursos.

Agradecimientos

El autor agradece a Raquel Behar-Lagares, María Angeles Jiménez Souza, Amanda Fernández Rodríguez, Rafael Blancas Gómez-Casero y Remo Suppi por su aportación y ayuda a la hora de realizar este trabajo.

Bibliografía

[1] World Health Organitation, «Covid-19 cases». Accedido: 28 de mayo de 2025. [En línea]. Disponible en: https://data.who.int/dashboards/covid19/cases?m49=001

[2] World Health Organization, «Global excess deaths associated with COVID-19 (modelled estimates)». Accedido: 28 de mayo de 2025. [En línea]. Disponible en: https://www.who.int/data/sets/global-excess-deaths-associated-with-covid-19-modelled-estimates

[3] R. Flisiak et al., «Variability in the Clinical Course of COVID-19 in a Retrospective Analysis of a Large Real-World Database», Viruses, vol. 15, n.o 1, ene. 2023, doi: 10.3390/V15010149,.

[4] R. Chen et al., «Prediction of prognosis in COVID-19 patients using machine learning: A systematic review and meta-analysis», Int J Med Inform, vol. 177, sep. 2023, doi: 10.1016/j.ijmedinf.2023.105151.

[5] P. Wu et al., «The trans-omics landscape of COVID-19», Nat Commun, vol. 12, n.o 1, dic. 2021, doi: 10.1038/s41467-021-24482-1.

[6] C. Hu et al., «Early prediction of mortality risk among patients with severe COVID-19, using machine learning», Int J Epidemiol, vol. 49, n.o 6, pp. 1918-1929, dic. 2020, doi: 10.1093/ije/dyaa171.

[7] R. Laguna-Goya et al., «IL-6–based mortality risk model for hospitalized patients with COVID-19», Journal of Allergy and Clinical Immunology, vol. 146, n.o 4, pp. 799-807.e9, oct. 2020, doi: 10.1016/j.jaci.2020.07.009.

[8] K. A. Overmyer et al., «Large-Scale Multi-omic Analysis of COVID-19 Severity», Cell Syst, vol. 12, n.o 1, pp. 23-40.e7, ene. 2021, doi: 10.1016/j.cels.2020.10.003.

[9] M. J. Pons et al., «Cytokine Profiles Associated With Worse Prognosis in a Hospitalized Peruvian COVID-19 Cohort», Front Immunol, vol. 12, sep. 2021, doi: 10.3389/fimmu.2021.700921.

[10] V. R. Richard et al., «Early Prediction of COVID-19 Patient Survival by Targeted Plasma Multi-Omics and Machine Learning», Molecular and Cellular Proteomics, vol. 21, n.o 10, oct. 2022, doi: 10.1016/j.mcpro.2022.100277.

[11] N. C. Soares, A. Hussein, J. S. Muhammad, M. H. Semreen, G. El Ghazali, y M. Hamad, «Plasma metabolomics profiling identifies new predictive biomarkers for disease severity in COVID-19 patients», PLoS One, vol. 18, n.o 8 August, ago. 2023, doi: 10.1371/journal.pone.0289738.

[12] E. H. Abdelaziz, R. Ismail, M. S. Mabrouk, y E. Amin, «Multi-omics data integration and analysis pipeline for precision medicine: Systematic review.», Comput Biol Chem, vol. 113, p. 108254, dic. 2024, doi: 10.1016/j.compbiolchem.2024.108254.

[13] J. Krumsiek, J. Bartel, y F. J. Theis, «Computational approaches for systems metabolomics», Curr Opin Biotechnol, vol. 39, pp. 198-206, jun. 2016, doi: 10.1016/j.copbio.2016.04.009.

[14] K. M. Kuo, P. C. Talley, y C. S. Chang, «The accuracy of machine learning approaches using non-image data for the prediction of COVID-19: A meta-analysis», Int J Med Inform, vol. 164, ago. 2022, doi: 10.1016/j.ijmedinf.2022.104791.

[15] K. Liu et al., «A systematic meta-analysis of immune signatures in patients with COVID-19», Rev Med Virol, vol. 31, n.o 4, jul. 2021, doi: 10.1002/rmv.2195.

[16] A. Tárnok, «Machine Learning, COVID-19 (2019-nCoV), and multi-OMICS», Cytometry Part A, vol. 97, n.o 3, pp. 215-216, mar. 2020, doi: 10.1002/cyto.a.23990.

[17] D. Feldner-Busztin et al., «Dealing with dimensionality: the application of machine learning to multi-omics data», Bioinformatics, vol. 39, n.o 2, feb. 2023, doi: 10.1093/bioinformatics/btad021.

[18] P. C. Austin y E. W. Steyerberg, «Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models», Stat Methods Med Res, vol. 26, n.o 2, pp. 796-808, abr. 2017, doi: 10.1177/0962280214558972,.

[19] V. M. T. De Jong et al., «Clinical prediction models for mortality in patients with covid-19: external validation and individual participant data meta-analysis», The BMJ, vol. 378, 2022, doi: 10.1136/BMJ-2021-069881,.

[20] L. De Rop et al., «Accuracy of routine laboratory tests to predict mortality and deterioration to severe or critical COVID-19 in people with SARS-CoV-2», Cochrane Database of Systematic Reviews, vol. 2024, n.o 8, ago. 2024, doi: 10.1002/14651858.CD015050.PUB2,.

[21] C. Buttia et al., «Prognostic models in COVID-19 infection that predict severity: a systematic review», Eur J Epidemiol, vol. 38, n.o 4, pp. 355-372, abr. 2023, doi: 10.1007/s10654-023-00973-x.

[22] F. P. Havers et al., «COVID-19-Associated Hospitalizations Among Vaccinated and Unvaccinated Adults 18 Years or Older in 13 US States, January 2021 to April 2022», JAMA Intern Med, vol. 182, n.o 10, p. 1071, oct. 2022, doi: 10.1001/JAMAINTERNMED.2022.4299.