Uso de *machine learning* para el análisis de pacientes hospitalizados con COVID-19 durante la entrada de la variante Omicron a Colombia como una herramienta para la toma de decisiones en salud pública.

Jhonnatan Reales González[a], Sabrina Bonill[b].

a Estudiante de Especialización en Estadística Aplicada; jdrealesg@libertadores.edu.co.
b Estudiante de Especialización en Estadística Aplicada; isbonillv@libertadores.edu.co

Resumen

Desde la emergencia de SARS-CoV-2 en 2019, múltiples linajes han sido reportados a nivel mundial. La aparición de nuevas mutaciones ha llevado a que la OMS denomine nuevas variantes como variantes de preocupación (*VOC*) o de interés (*VOI*) y ha hecho un llamado a priorizar la vigilancia y análisis del efecto de estas variantes en las distintas poblaciones. En colombia, la variante Omicron desplazó a la *VOC* Delta desde finales de diciembre, 2021 y a principios del 2022 se observó un incremento en los casos de hospitalización con casos de COVID-19. Con el objetivo de caracterizar estos pacientes, se realizó un censo de pacientes hospitalizados por esta causa en tres departamentos del país. Pruebas moleculares fueron realizadas e información epidemiológica, de hospitalización y antecedentes vacunales fueron recolectadas y análizadas mediante técnicas de machine learning. Se encontró que, en la etapa de post-vacunación, los principales factores de riesgo fueron edades mayores a 60 años, no tener inmunidad previa y/o contar con una vacunación primaria mayor a 200 días sin ninguna dosis de refuerzo. Las variantes de preocupación aquí estudiadas no se asociaron con el desenlace de la enfermedad. Nuestros resultados demuestran la utilidad del ML para la caracterización de las

poblaciones afectada tras la emergencia de nuevas variantes de preocupación, necesaria para la toma de desiciones en salud pública de una región.

INTRODUCCIÓN

El nuevo coronavirus, SARS-CoV-2, emergió a finales del 2019 y rápidamente se propagó alrededor de todo el mundo, siendo considerado como una pandemia poco tiempo después por la OMS. Hasta el 14 de junio de 2022, 535'607.511 han sido reportados en todo el mundo, con 6,310,254 casos fatales (https://coronavirus.jhu.edu/map.html).

Desde el reporte de los primeros casos, una gran diversidad genética de SARS-CoV-2 ha sido identificada debido a su alta capacidad de transmisión y aislamiento geográfico (Zhou et al., 2021). La emergencia de nuevas variantes es el resultado de un proceso natural que ocurre cuando los virus tienen elevadas tasas de replicación (dos Santos, 2021). Asi mismo, la adquisición de múltiples mutaciones en el genoma del virus llevó a la clasificación de genotipos divergentes como Variantes de Preocupación (*VOC*) y Variantes de Interés (*VOI*) por la OMS con el objetivo de priorizar su vigilancia.

Aunque desde el 2020 inició la etapa de vacunación contra COVID-10 y 11,552,796,345 de dosis han sido aplicadas en el mundo, las variantes de preocupación han mostrado producir una disminución en la eficacia de medidas sociales, de diagnóstico y de vacunas o tratamientos disponibles (OMS, 2022). Debido a esto, la OMS propuso priorizar el seguimiendo de estas nuevas variantes y realizar estudios de análisis pertinentes que permitan conocer mejor los efectos de nuevas variantes en las caracteristicas epidemiológicas de la COVID-19 en distintas regiones del mundo (OMS, 2022). Análisis de grandes muestreos son necesarios para determinar si estos nuevos clados virales tienen una influencia significativa en el desenlace final de los pacientes (Nakamichi et al., 2021).

La pandemia por SARS-CoV-2 también ha generado una gran cantidad de datos en todo el mundo, conllevando a la urgente necesidad de desarrollar respuestas efectivas mediantes el uso de *Machine Learning* (ML). Enfoques diferentes a procedimientos clínicos y estudios epidemiológicos tradicionales, como el ML, data mining y otras técnicas de inteligencia artificial pueden apoyar el análisis de diagnóstico, prognosis y terapias para la enorme cantidad de datos colectados en casos de infección por SARS-CoV-2 (Mottaqi, Mohammadipanah, & Sajedi, 2021). Las técnicas de ML tienen la ventaja que permiten encontrar patrones subyacentes en los datos que no pueden ser encontrados con análisis tradicionales y pueden ser divididas en aprendizaje supervisado y no supervisado. Los algoritmos supervisados pueden aprender de casos del pasado para predecir futuros eventos; en estos, los datos son divididos en categorias etiquetadas para entrenar al modelo, y la máquina aprende a identificar esas etiquetas de clase en la variable respuesta en grupos de datos distintos a los del entrenamiento. Por su parte, el objetivo de los algoritmos no supervisados es modelar y usar información que no está clasificada o etiquetada (Mottaqi et al., 2021) para encontrar patrones escodidos y categorias naturales de los datos (Biamonte et al., 2017; Mottaqi et al., 2021). Ambos tipos de algoritmos de ML han sido funcioanles para análisis de información relacionada a SARS-CoV-2.

Teniendo en cuenta lo anterior, el objetivo de este trabajo fue caracterizar pacientes hospitalizados con COVID-19 tras la entrada de la *VOC* Omicron a Colombia usando técnicas de machine learning (ML) para ayudar a la toma de desciones en salud pública.

ESTADO DEL ARTE

En uso del machine learning (ML) en las ciencias biológicas ha sido fundamental a para analizar la gran cantidad de datos producidos en laboratorios o en muestreos realizados con fines de toma de decisiones en salud pública. Asi mismo, se ha hecho una herramienta muy importante durante la pandemia de SARS-CoV-2 debido al gran volumen de información obtenida alrededor de todo el mundo. Shoukat *et al*., (2021) usaron ML para analizar datos de secuencias de células inmune de pacientes recuperados de infecciones causadas por SARS-CoV-2 que habían tenido una baja severidad de la enfermedad y de individuos que no habian sido infectados. Inicialmente, usaron un metodo de reducción de dimensiones (PCA) de los datos obtenidos y una combinación de los primeros componentes principales fueron evaluados para determinar la mejor combinación que permitiera separar pacientes recuperados de los no infectados. Este método permitió la correcta clasificación de los pacientes usando secuencias de las células T. Sin embargo, el modelo no fue capaz de separar pacientes recuperados de aquellos no infectados previamente basandose en los datos de los receptores de las células B, sugiriendo que las respuesta de las celulas T es más duradera contra SARS-CoV-2. Estos resultados demuestran la importacia de implementar modelos de machine learning en la caracterización de pacientes infectados por este patogeno. Sin embargo, los autores resaltan la necesidad de usar este efoque en largas cohortes de pacientes. Asimismo, es importante aplicar estos modelos a nivel poblacional para tener un mejor entendimiento de la pandemia con la emergencia de nuevas variantes.

Con la emergencia de nuevas variantes de preocupacion de SARS-CoV-2, se hace necesario evaluar distitnos tipos de análisis que permitan un mejor entendimiento

de los pacientes afectados de forma rápida para encontrar patrones subyacentes, incluso a poblaciones ya inmunizadas. Para comparar la infección de pacientes vacunados en distintos distritos de Jakarta, Indonesia, por las variantes delta y omicron, (Desy et al., 2022) compararon diferentes métodos de clusterización. Los autores usaron el Dunn Index y Hubert Index para determinar el mejor número de *clusters*. Asi mismo, compararon los métodos de Silhouette y de Davies Bouldien para encontrar el mejor método de clusterización entre los métodos Fuzzy C-means, K-means y PAM (*Partition Around Medioids*).

El método Silhouette es una medida estadística que sirve para encontrar el número de clústers óptimo y provee una evaluación de la validez de la clusterización. Basicamente, en este método cada clúster es denominado silueta y muestra cual objeto encaja bien en cada clúster y cuales están entre distintos clústers. Luego las siluetas son combinadas en un solo gráfico y permite una apreciación de la configuración de los datos y de la calidad de los clusters (Rousseeuw, 1987). Por su parte, el método de *K-means* es un método iterativo de clusterización por medio de reasignación, donde se selecciona un número de puntos igual al número de clusters deseados y luego cada observación es asociada a su centro más cercado para crear clusters temporales, esto se repite hasta que el centro no se mueve y se convierte en el nuevo cluster (Hartigan & Wong, 1979).

Los resultados de (Desy et al., 2022) mostraron que el mejor método de clusterización para analizar pacientes infectados por las variantes Delta y Omicron es el *K-means*, el cual mostró los valores más altos de Silhouette y más bajos de Davies Boulding comparado con los otros métodos. Esto indica que el método de *K-means* es apropiado para implementar clusterización con las variantes Delta y Omicron.

De manera similar (Virgantari & Faridhan, 2020) usaron el metodo de K-means en la clusterización de casos de COVID-19 en provincias de Indonesia y los analizaron con distitntos números de clusters predefinidos para determinar el mejor número de clusters a usar. Actualmente, existen métodos de clusterización jerarquicos y no jerarquicos, *K-means* pertenece a estos últimos debido a que necesita el número de clusters deseados antes del análisis (Hartigan & Wong, 1979). Ademas, *K-means* es el método más popular en la sub-categoría de particionamiento. Debido a que este método es no jerarquito, no hay necesidad de hacer un dendograma para la clusterización según estos autores. Los resultados obtenidos mostraron que este método permitió la clasificación exitosa de casos de COVID-19 por provincias, lo

que es util para advertir sobre la propagación de esta enfermedad y para la toma de decisiones óptimas durante la pandemia.

En cuanto al correcto número de clusters a escoger, Johnson & Wichern, (2007) recomiendan tener en cuenta algunos puntos importantes durante este paso. Los clusters deben estar muy bien diferenciados; si hay outliers, se generaría al menos un grupo con objetos no tan similares y, por último, forzar los datos a un número de grupos pre-definidos puede dar lugar a la creación de clusters sin sentido. Esto es importante y debe ser considerado a la hora de realizar los análisis necesarios para la clusterización.

El particionamiento mediante el método *K-means* tambien ha sido usado para la caracterización de mutaciones de SARS-CoV-2 en Estados Unidos, lo que es esencial para enteder el comportamiento de este patógeno en un pais. Además de otro tipo de analisis epidemiológicos y genómicos, (Wang et al., 2020) usaron este método para el análisis de mutaciones específicas en el genoma del virus. Además, el análisis fue complementado con el uso del método de codo (Elbow) para analizar el óptimo número de subtipos de de variantes SNP de SARS-CoV-2. Los resultados permitieron agrupar las cepas del virus de manera correcta en 4 clusters caracterizados por distintas mutaciones de origen diverso. Adicionalmente, los resultados obtenidos sugieren que el sistema inmune femenino es más activo que el de los hombres al responder a infecciones por SARS-CoV-2 e identificaron 4 sub-cepas más contagiosas. Estos resultados demuestran que la aplicación del método *K-means* complementado con el método del codo son altamente útiles a la hora de analizar información relacionada a SARS-CoV-2 para una mejor comprensión del mismo y poder ayudar a tomar mejores decisiones en salud pública.

Los métodos de clusterización pueden ser grandemente complementados por otros modelos de machine learning supervisados para analizar pacientes infectados con SARS-CoV-2. (Nakamichi et al., 2021) buscaron determinar si las secuencias de variantes de SARS-CoV-2 están asociadas con el estado final de los pacientes infectados. Las secuencias de genoma total se obtuvieron a partir de muestras de pacientes de un sistema médico durante un mes y medio en el 2020. La información demográfica, características clínicas estado final y datos de hospitalización de los pacientes fue obtenida. Modelos estadísticos y de machine learning fueron aplicados para determinar si variantes genética virales se asocian con la hospitalización o muerte de los pacientes. Los autores consideraron cuatro formar para seleccionar las mejores variables a incluir en los modelos. Seguidamente, evaluaron distintos

modelos de machine learning donde se incluyeron combinaciones de las distintas variables disponibles (demográficas, clínicas, linajes y datos genéticos). Cada modelo fue corrido usando el método de validación cruzada y el ajuste del modelo se logró con 10.000 conjuntos aleatorios de hiperparametros para 4 model (AdaBoost, Extra Trees, Gradient Boosting, Random Forest) de la librería scikit-learn (v0.22.2). En este estudio, los métodos de clusterización mostraron 2 clados virales mayores, que se distinguieron por 12 polimorfismos en 5 genes y se observó una mayor tendencia de hospitalizacion para pacientes con infecciones del cluster 2. Y la comparación de los modelos de machine learning mostraron que el modelo de Random forest fue ligeramente superior a los demás evaluados. De igual forma, el modelo que usó solo la información demográfica y las co-morbilidades obtuvo el mejor valor del área bajo la curva (0.93) para predecir la hospitalización. Otras variables como el linaje viral o información genómica no mejoraron la capacidad predictiva del modelo.

Los estudios realizados a la fecha, muestran que el método de K-means es adecuado para realizar análisis de clusterización al evaluar información obtenida de pacientes infectados por SARS-CoV-2. Asimismo, el algoritmo de machine learning Ramdon forest permite crear modelos de predicción más precisos cuando se trata de investigar múltiples variables dentro de un set de datos en el contexto de la actual pandemia.

METODOLOGÍA

Consideraciones Éticas

La información de todos los pacientes fue anonimizada para la protección de su identidad. La información utilizada en el presente trabajo, asi como los resultados obtenidos, pertenece al Instituto Nacional de Salud como resultado de la vigilancia epidemiologica y genómica. El presente documento se entrega como requisito académico a la Fundación Universitaria Los Libertadores, pero se debe garantizar la privacidad de los resultados aquí mostrados.

Obtención de Muestras

Durante la tercera semana epidemiológica del año 2022 se observó un aumento de casos de hospitalización de pacientes con COVID-19 en Colombia tras la entrada de la variante de preocupación Ómicron al país a finales del año 2021. Para analizar estos casos, se realizó un censo de pacientes que habían sido ingresados a instituciones hospitalarias tras la confirmación de COVID-19 mediante RT-PCR y cuadro clínico. El resultado de la RT-PCR en tiempo real se mide por el valor de Ct (cyclce threshold), este valor es un indicativo de la carga viral del paciente al momento del muestreo. Se realizó secuenciación genómica para obtener linaje o variante causante de la infección.

Obtención de Información

Información epidemiológica de los pacientes (edad, género, desenlace clínico, reinfección, días de hospitalización y antecedentes clínicos) fue recolectada posteriormente mediante bases de datos nacionales que hacen parte del Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA). Adicionalemente, el historial

de vacunación de los pacientes fue recolectado a partir del sistema PAIWEB del Ministerio de Salud y Protección Social. Las variables para las que no se encontró documentación de los pacientes, se codificaron como "sin dato" y no se tuvieron en cuenta para los análisis multivariados o pruebas estadísticas.

Análisis Estadísticos

- **Estadística descriptiva**

Se realizó un análisis exploratorio de los datos obtenidos de los pacientes censados. Las caracteristicas epidemiológicas y vacunales fueron resumidas usando estadística descriptiva como frecuencias absolutas y relativas, y se realizó una pruba de asociación de Pearson (X^2) para las distintas variables estudiadas con las variantes mayormente detectadas (Omicron y Delta) como variable de interés. Las variables incluidas en los análisis multivariados y de machine learning se escogieron mediante el criterio de Akaike (AIC) y según criterio de los investigadores tras el análisis exploratorio inicial.

- **Análisis multivariado**

Teniendo en cuenta la importancia de la variante causante de la infección, para los posteriores análisis se escogió una sub-muestra del total de datos que incluyó todos los casos donde se obtuvo el linaje del virus. Y de estos, se seleccionaron sólo los casos completos, con excepción de la variable "Dias de hospitalización", la cual presentó un 8.5% de datos faltantes (*missing values*). Estos datos fueron imputados con el valor de la mediana del departamento del cual provenía el paciente. Se realizó un análisis descriptivo multivariante y los supuestos del análisis multivariado (normalidad, homocedasticidad y linealidad) fueron verificados.

Con el objetivo de caracterizar mejor estos pacientes e intentando buscar patrones dentro de este grupo, la información obtenida fue analizada usando clusterización mediante el método de *K-Means* teniendo en cuenta variables numéricas de importancia relacionadas con el tiempo de vacunación y días de hospitalización hasta la fecha del censo. El número óptimo de clusters fue previamente obtenido mediante tres distintos métodos, el método del codo, Silhouette y Gap statistics. Para la clusterización las variables fueron estandarizadas (escaladas) con el objetivo de hacerlas comparables. Las caracteristicas de los pacientes pertenecientes a cada uno de los *clusters* fueron descritas y analizadas posteriormente.

- **Machine Learning**

Nuestra variable respuesta de interés fue el desenlace del paciente. Sin embargo, la base final presentó un total de 341 y 66 pacientes vivos y fallecidos, respectivamente. Con el objetivo de no condicionar los modelos de machine learning propuestos, se realizó un balanceo de datos para tener el mismo número de observaciones de pacientes fallecidos y vivos mediante el método de sobremuestreo Random Over-Sample (ROS). Durante el preprosesamiento, las variables en el data set final fueron dummisadas y escaladas con el objetivo de obtener modelos con mejores resultados de predicción.

Se emplearon distintos modelos de ML de la librería Scikit-learn (version 1.1.1). Cada modelo fue construido con nuestros datos teniendo en cuenta las variables demográficas, vacunales y clínicas. El dataset fue dividido en grupos de entrenamiento y de prueba con el 70% y 30% de los datos, respectivamente. Las métricas usadas para evaluar el rendimiento de cada modelo fueron el *accuracy* (porcentaje de observaciones correctamente clasificadas respecto al total de predicciones) (Amat, 2016) y el área bajo curva ROC calculada tras la validación cruzada realizada con los datos de prueba. Las curvas ROC muestran la relación que existe entre la tasa de verdaderos positivos y la tasa de falsos positivos del modelo (MathWorks, 2022). Los modelos con un *accuracy* obtenido de 80% o mayor, fueron validados mediante el método de validación cruzada (cross-validation) con 5 iteraciones (5 pliegues).

RESULTADOS

Análisis exploratorios de los datos

Inicialmente, se realizó un análisis exploratorio de todas las variables obtenidas y, en general, los valores faltantes estuvieron entre un 4 y 10% para las distintas variables (fig. 1). La variables con más datos faltantes fue la variante debido a que sólo se secuenció el 40% de los casos totales (Tabla S1).

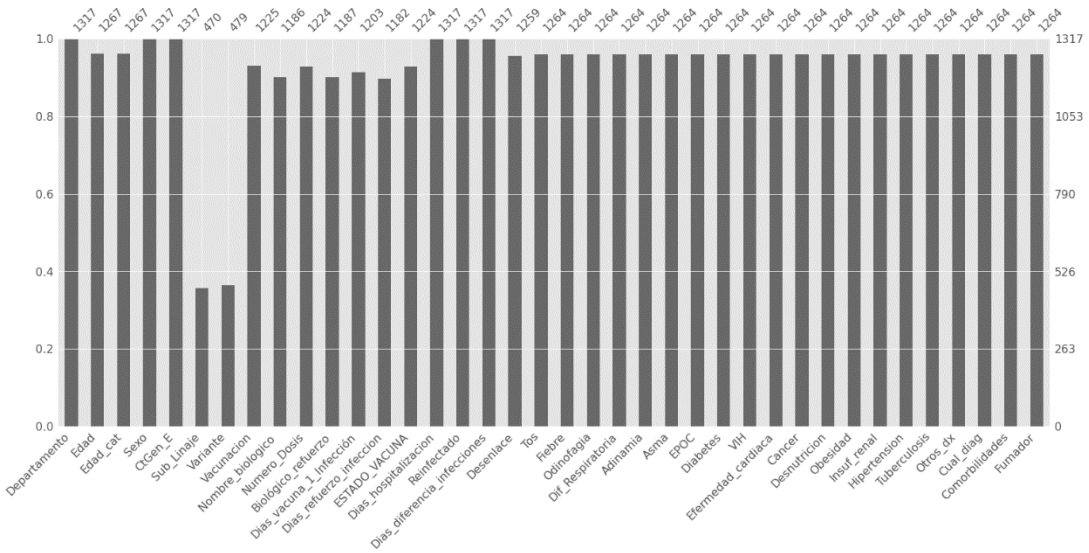

Figura 1. Gráfico de valores faltantes en la base de datos inicial. Las barras indican el número y porcentaje de valores completos por cada variable.

La distribución de cada variable fue graficada, la edad mostró una mayor frecuencia en edades mayores a 60 años, el valor de Ct del gen E mostró que el valor más frecuente fue de aproximadamente 25 y los días de hospitalización estuvieron entre 1 y 60, con mayor concentración de datos durante los primeros 10 días. Por su parte,

el tiempo entre vacunación y una posterior infección presentó mayor cantidad de datos entre 200 y 300 días, el cero indica pacientes que no se habían vacunado. Debido a que sólo unos pocos pacientes tenían dosis de refuerzo o habían sido infectados por este virus previamente, se observó una gran concentración de datos con cero días para ambas variables (fig. 2). Entre las variables numéricas tambien se realizó un análsis de correlación para observar el comportamiennto y una posible colinealidad. La figura 3 muestra que no hubo una alta relación entre las distintas variables, por lo que inicialmente no se descartó ninguna de estas para posteriores análisis. Esta información se corroboró con una tabla de correlaciónes (Tabla S2) y mediante un gráfico de mapa de calor (fig. 4).

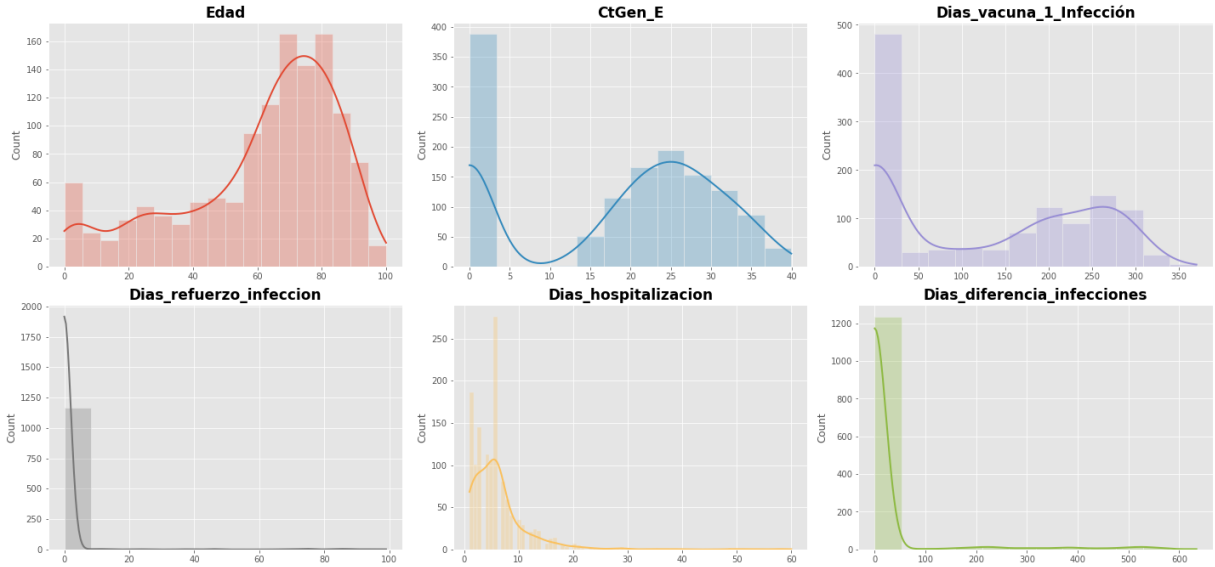

Figura 2. Distribución de variables numéricas medidas en este estudio.

Figura 3. Dispersión de datos entre distintas variables numéricas.

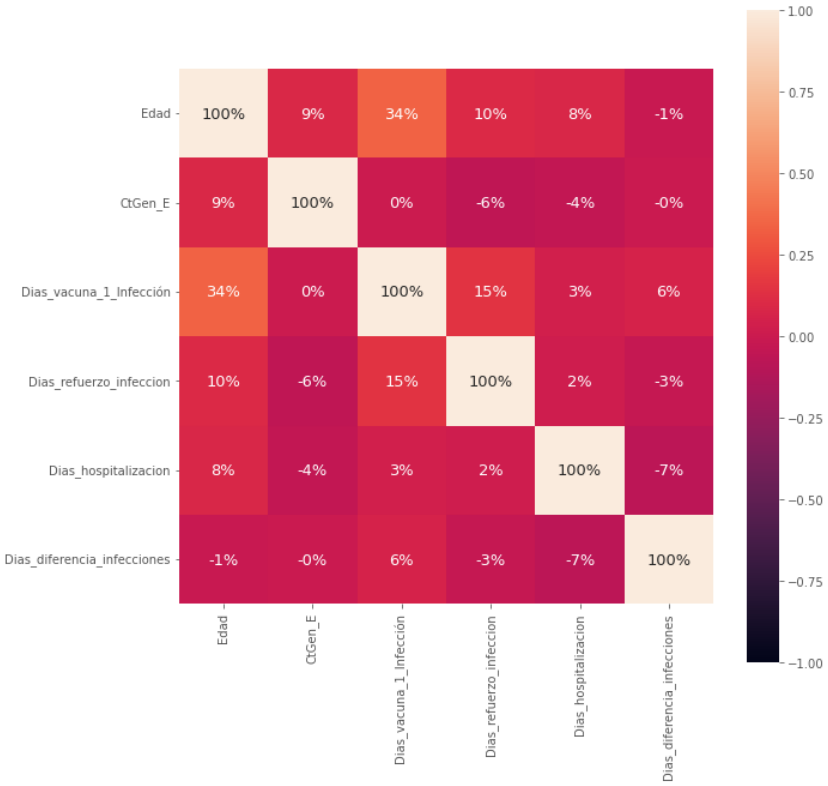

Figura 4. Mapa de calor con los grados de correlación entre las distitntas variables numéricas.

Características de los pacientes

Entre el 16 y 20 de enero de 2022, se recolectaron 1317 muestras de pacientes internados las distintas instituciones hospitalarias de tres departamentos principales de Colombia (Valle del Cauca, Atlántico y Antioquia). Tras realizar la detección molecular se encontró que, para el momento del muestreo, el 70% (929/1317) de los pacientes aún permanecían positivos para SARS-CoV-2 mediante RT-PCR en tiempo real. El resto de los pacientes aún permanecía hospitalizados a pesar de arrojar un resultado negativo.

Del total de pacientes hospitalizados, el 52,6 % (577/1317) fue hombres, con una mediana de edad de 68 años (IQR=48 -79 años). De manera general, más de la mitad de los pacientes en los tres departamentos tenían más de 60 años y Atlántico presentó mayor proporción de casos en personas jóvenes menores a 18 años (14,3%) en comparación con los otros dos departamentos ($p<0.05$). La información del estado final de los pacientes se actualizó dos meses después del estudio y se obtuvo que el 15% (197/1317) de los pacientes fallecieron. Sin embargo, los días de hospitalización

fueron contados hasta momento del estudio. Para los pacientes de Antioquia y Valle del Cauca el tiempo de hospitalización tuvo una mediana de 5 y 6 días respectivamente, muy por encima de la mediana de tiempo de hospitalización en Atlántico con una mediana de 1. Curiosamente, en los primeros dos departamentos se encontró una mayor proporción de la variante Delta, mientras que en Atlántico solo se identificó un caso (tabla 1). Antioquia fue el departamento que más presentó casos de reinfección entre los hospitalizados con un 12,4%.

De los casos positivos al momento del muestreo, el 52% (479/929) pudo ser secuenciado exitosamente y obtener el linaje causante de la infección. Se encontró que el 82,3% de estos casos fue producido por la variante Omicron, seguido por la variante Delta (16,9%) y se encontró un caso de la variante Mu.

Los síntomas con mayor frecuencia fueron tos, fiebre y adinamia, y las comorbilidades o enfermedades con mayor presencia entre los paceintes fue hipertensión, diabetes y EPOC. El 3,6% de los pacientes presentaba antecedentes como fumador (tabla 1).

Tabla 1. Caracteristicas epidemiológicas de los pacientes ingresados en el estudio.

Variable	Antioquia n= 364	Atlántico n= 239	Valle del Cauca n= 714	Total n= 1317	*p*-valor
Edad (años), mediana (IQR)	68 (50-78)	67 (41,5-79)	68 (50-79)	68 (48-79)	0,45
Grupo de edad					
0-9	20 (5,5)	25 (10,5)	25 (3,5)	70 (5,3)	
10-17	5 (1,4)	9 (3,8)	7 (1)	21 (1,6)	
18-39	47 (12,9)	24 (10)	73 (10,2)	144 (10,9)	
40-59	51 (14)	35 (14,6)	103 (14,4)	189 (14,4)	
> 60	241 (66,2)	143 (59,8)	403 (56,4)	787 (59,8)	<0.05
ND	0 (0)	3 (1,3)	103 (14,4)	106 (8)	
Género					
Femenino	152 (41,8)	110 (46)	315 (44,1)	577 (43,8)	
Masculino	212 (58,2)	129 (54)	399 (55,9)	740 (56,2)	0,57
ND	26 (7,1)	32 (13,4)	34 (4,8)	92 (7)	
Estado Final					
Vivo	283 (77,7)	196 (82)	583 (81,7)	1062 (80,6)	
Muerto	81 (22,3)	43 (18)	73 (10,2)	197 (15)	0,64
ND	0 (0)	0 (0)	58 (8,1)	58 (4,4)	

Reinfección					
Si	45 (12,4)	16 (6,7)	20 (2,8)	81 (6,2)	
No	319 (87,6)	223 (93,3)	694 (97,2)	1236 (93,8)	0,15
Días de Hospitalización					
Mediana (IQR)	5 (3-8)	1 (1-3)	6 (4-9)	5 (2-8)	<0.05
Linaje secuenciado					
Mu	0 (0)	1 (0,7)	0 (0)	1 (0,2)	
Delta	21 (22,8)	1 (0,7)	59 (23,8)	81 (16,9)	
Omicron	71 (77,2)	137 (98,6)	189 (76,2)	397 (82,8)	0,37
Tos					
No	128 (35,2)	80 (33,5)	242 (33,9)	450 (34,2)	
Si	232 (63,7)	141 (59)	441 (61,8)	814 (61,8)	
ND	4 (1,1)	18 (7,5)	31 (4,3)	53 (4)	<0.05
Fiebre					
No	193 (53)	117 (49)	381 (53,4)	691 (52,5)	
Si	167 (45,9)	104 (43,5)	302 (42,3)	573 (43,5)	
ND	4 (1,1)	18 (7,5)	31 (4,3)	53 (4)	<0.05
Odinofagia					
No	286 (78,6)	173 (72,4)	576 (80,7)	1035 (78,6)	
Si	74 (20,3)	48 (20,1)	107 (15)	229 (17,4)	
ND	4 (1,1)	18 (7,5)	31 (4,3)	53 (4)	<0.05
Dif_Respiratoria					
No	212 (58,2)	144 (60,3)	415 (58,1)	771 (58,5)	
Si	148 (40,7)	77 (32,2)	268 (37,5)	493 (37,4)	
ND	4 (1,1)	18 (7,5)	31 (4,3)	53 (4)	<0.05
Adinamia					
No	145 (39,8)	155 (64,9)	404 (56,6)	704 (53,5)	
Si	215 (59,1)	66 (27,6)	279 (39,1)	560 (42,5)	
ND	4 (1,1)	18 (7,5)	31 (4,3)	53 (4)	<0.05
Asma					
No	351 (96,4)	214 (89,5)	676 (94,7)	1241 (94,2)	
Si	9 (2,5)	7 (2,9)	7 (1)	23 (1,7)	
ND	4 (1,1)	18 (7,5)	31 (4,3)	53 (4)	<0.05
EPOC					
No	301 (82,7)	210 (87,9)	620 (86,8)	1131 (85,9)	
Si	59 (16,2)	11 (4,6)	63 (8,8)	133 (10,1)	
ND	4 (1,1)	18 (7,5)	31 (4,3)	53 (4)	<0.05
Diabetes					
No	297 (81,6)	193 (80,8)	569 (79,7)	1059 (80,4)	
Si	63 (17,3)	28 (11,7)	114 (16)	205 (15,6)	
ND	4 (1,1)	18 (7,5)	31 (4,3)	53 (4)	<0.05
VIH					

No	354 (97,3)	218 (91,2)	680 (95,2)	1252 (95,1)	
Si	6 (1,6)	3 (1,3)	3 (0,4)	12 (0,9)	
ND	4 (1,1)	18 (7,5)	31 (4,3)	53 (4)	<0.05
Efermedad_cardiaca					
No	321 (88,2)	205 (85,8)	637 (89,2)	1163 (88,3)	
Si	39 (10,7)	16 (6,7)	46 (6,4)	101 (7,7)	
ND	4 (1,1)	18 (7,5)	31 (4,3)	53 (4)	<0.05
Cáncer					
No	349 (95,9)	215 (90)	642 (89,9)	1206 (91,6)	
Si	11 (3)	6 (2,5)	41 (5,7)	58 (4,4)	
ND	4 (1,1)	18 (7,5)	31 (4,3)	53 (4)	<0.05
Desnutrición					
No	356 (97,8)	221 (92,5)	673 (94,3)	1250 (94,9)	
Si	4 (1,1)	0 (0)	10 (1,4)	14 (1,1)	
ND	4 (1,1)	18 (7,5)	31 (4,3)	53 (4)	<0.05
Obesidad					
No	336 (92,3)	213 (89,1)	638 (89,4)	1187 (90,1)	
Si	24 (6,6)	8 (3,3)	45 (6,3)	77 (5,8)	
ND	4 (1,1)	18 (7,5)	31 (4,3)	53 (4)	<0.05
Insuf_renal					
No	341 (93,7)	213 (89,1)	647 (90,6)	1201 (91,2)	
Si	19 (5,2)	8 (3,3)	36 (5)	63 (4,8)	
ND	4 (1,1)	18 (7,5)	31 (4,3)	53 (4)	<0.05
Hipertensión					
No	237 (65,1)	172 (72)	467 (65,4)	876 (66,5)	
Si	123 (33,8)	49 (20,5)	216 (30,3)	388 (29,5)	
ND	4 (1,1)	18 (7,5)	31 (4,3)	53 (4)	<0.05
Tuberculosis					
No	357 (98,1)	221 (92,5)	680 (95,2)	1258 (95,5)	
Si	3 (0,8)	0 (0)	3 (0,4)	6 (0,5)	
ND	4 (1,1)	18 (7,5)	31 (4,3)	53 (4)	<0.05
Fumador					
No	333 (91,5)	219 (91,6)	665 (93,1)	1217 (92,4)	
Si	27 (7,4)	2 (0,8)	18 (2,5)	47 (3,6)	
ND	4 (1,1)	18 (7,5)	31 (4,3)	53 (4)	<0.05

ND = No dato

En cuanto a la variable de antecedente de vacunación, se identificó que el 57,9 % (762/1317) de los pacientes reportaba algún antecedente de vacunación para COVID-19, con una mediana de tiempo entre la primera dosis y la fecha de muestreo de 220 días (IQR= 170 – 272 días). La información del tiempo de vacunación y las dosis aplicadas fue utilizada para crear las categorías de vacunación "óptima",

"adecuada" y "no vacunados". En la primera categoría se incluyen los pacientes que presentan 3 o 2 dosis con menos de 180 días de aplicación, el 4.9% (64/1317) de los pacientes estaba en este grupo. Una vacunación adecuada se definió como aquella en que los pacientes tienen 2 dosis con más de 180 o 1 de Janssen y estuvo representa por el 45.4% (598/1317) de los casos. Los no vacunados se tomaron como aquellos que tienen solo una dosis o ninguna dosis y el 42.7% (562/1317) de los casos se encontraron en esta categoría (Tabla 2).

El 73.1% (557/762) de los pacientes vacunados tenía un esquema con 2 dosis y los biológicos con mayor porcentaje de aplicación fueron CoronaVac - Sinovac con 46.9% (357/762) y Pfizer con 21.9 % (167/762). Se identificó que el 8,4 % (64/762) de los vacunados tenía dosis de refuerzo, mayormente con CoronaVac – Sinovac.

Tabla 2. Antecedente vacunal de los pacientes hospitalizados por COVID-19.

Variable	Antioquia n= 364	Atlántico n= 239	Valle del Cauca n= 714	Total n= 1317	*p*-valor
Vacunación					
Si	230 (63,2)	129 (54)	403 (56,4)	762 (57,9)	
No	108 (29,7)	78 (32,6)	277 (38,8)	463 (35,2)	0,75
ND	26 (7,1)	32 (13,4)	34 (4,8)	92 (7)	
Estado de vacunación					
Optima	24 (6,6)	14 (5,9)	26 (3,6)	64 (4,9)	
Adecuada	174 (47,8)	96 (40,2)	328 (45,9)	598 (45,4)	
No vacunado	140 (38,5)	96 (40,2)	326 (45,7)	562 (42,7)	0,25
ND	26 (7,1)	33 (13,8)	34 (4,8)	93 (7,1)	
Dosis Aplicadas*					
1	48 (20,9)	26 (20,3)	66 (16,4)	140 (18,4)	
2	158 (68,7)	88 (36.7)	311 (77,2)	557 (73,1)	
3	24 (10,4)	14 (10,9)	26 (6,5)	64 (8,4)	0,36
Tipo de vacuna*					
AstraZeneca	51 (22,2)	17 (13,2)	54 (13,4)	122 (16,01)	
Johnson & Johnson - Janssen	16 (7)	8 (6,2)	17 (4,2)	41 (5,4)	
ARNm-1273 - Moderna	12 (5,21)	11 (8,5)	13 (3,2)	36 (4,7)	
BioNTech, Pfizer	49 (21,3)	31 (24)	87 (21,6)	167 (21,9)	0,68
CoronaVac - Sinovac	94 (40,9)	55 (42,6)	208 (51,6)	357 (46,9)	
ND	8 (3,3)	7 (5,4)	24 (1,7)	39 (5,11)	
Tipo de vacuna de refuerzo**					
AstraZeneca	2 (8,33)	1 (7,14)	3 (11,5)	6 (9,38)	
ARNm-1273 - Moderna	2 (8,33)	0 (0)	6 (23,1)	8 (10,9)	
BioNTech, Pfizer	4 (16,7)	0 (0)	6 (23,1)	10 (15,6)	

CoronaVac - Sinovac	11 (45,8)	2 (14,3)	9 (34,6)	22 (34,4)	
ND	5 (20,8)	11 (78,6)	2 (7,7)	19 (29,7)	0,78

*El porcentaje es calculado entre los pacientes vacunados.

** El porcentaje es calculado entre pacientes con 3 dosis.

Cuando se analizaron los pacientes infectados por las variantes Delta y Omicron, se encontró que la variante Delta afectó a pacientes con rango de edades más jóvenes, mientras que Omicron tuvo mayor proporción en pacientes infectados mayores a 60 años ($p < 0.05$). No se encontró diferencia en las demás variables analizadas entre las variantes de interés (tablas S3, S4, S5).

Al estratificar a los pacientes por el desenlace de la infección (vivos, fallecidos), se encontró que los fallecidos tenían mayor número de días de hospitalización y una mediana de edad más alta que los pacientes vivos. Asimismo, hubo mayor proporción de pacientes que no habían contraído una infección previa entre los fallecidos en comparación con el grupo de los vivos (tabla S6). Contrariamente, no hubo diferencia del antecedente vacunal (vacunación óptima, adecuada y no vacunados) entre vivos y fallecidos (tabla S7).

Análisis multivariados

Se realizó un análisis descriptivo multivariado, analizando los vectores de medias de variables numéricas, los resultados hacen parte de la tabla 1. El gráfico de Andrews es una herramienta importante para analizar problemas multivariados y se recomienda para presentar las observaciones multivariadas (Khattree & Naik, 2002). En la figura 4 se muestra el gráfico de Andrews obtenido para nuestros datos. Se observa que la mayoría de los pacientes estudiados presentaron características similares en conjunto, dando lugar a un patrón que se observa en curvas de una misma clase; pocos pacientes fueron representados con curvas de distintas formas.

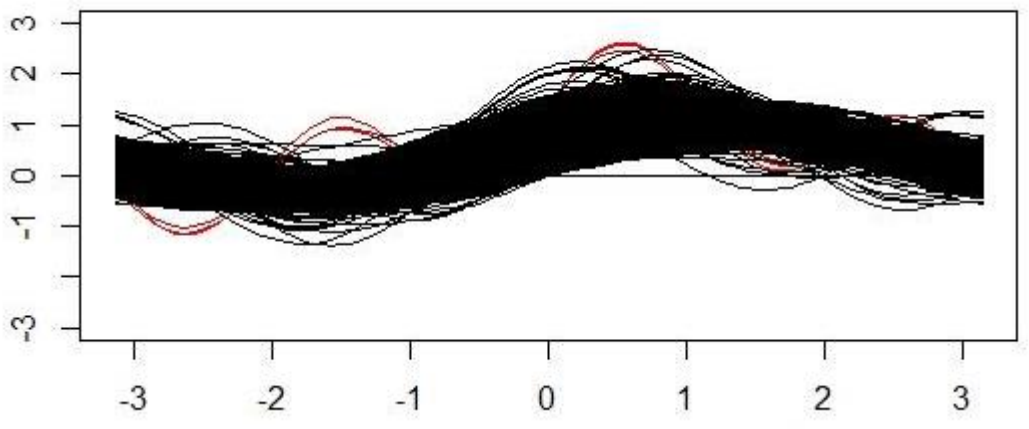

Figura 4. Gráfico de Andrews de la información multivariada de los pacientes analizados.

Se cumplieron los supuestos de análisis multivariados, con excepción de la normalidad multivariada de los datos (Shapiro-Wilk, $p < 0.05$). Sin embargo, teniendo en cuenta la cantdad de datos anlizados y el teorema del límite central, se asume que los análisis no se ven afectados por dicha falta de normalidad. Con el análisis de homogeneidad de matrices de varianza-covarianza se realizó una pruba M de Box y se observó una mayor variabilidad en el grupo de vivos cuando se comparó con el grupo de fallecidos ($p < 0.05$). Debido a que la prueba anterior puede verse afectada por la falta de normalidad de los datos, también se realizó la prueba T de Hotteling y se obtuvieron resultados iguales, por lo que se rechaza la hipótesis nula de la prueba ($p < 0.05$).

Clusterización mediante método *K-means*

Para complementar los análisis multivariados, se realizó una clusterización de los datos utilizando el método de *k-means.* El número óptimo de *clusters* fue estimado mediante tres métodos diferentes. Con el método del codo (Elbow) se busca seleccionar la cantidad ideal de grupos a partir de la optimización de la suma de los cuadrados dentro de los clusters y el número de clusters se escoge con el mayor punto de inflexión en la curva del gráfico. Con los otros métodos el número óptimo de grupos aparece automaticamente en el gráfico. El método del codo obtuvo como grupos óptimos un k=4 (fig. 5-A), el método de silueta obtuvo un k= 9 (fig. 5-B) y el método de Gap Statistics un k= 2 (fig. 5-C).

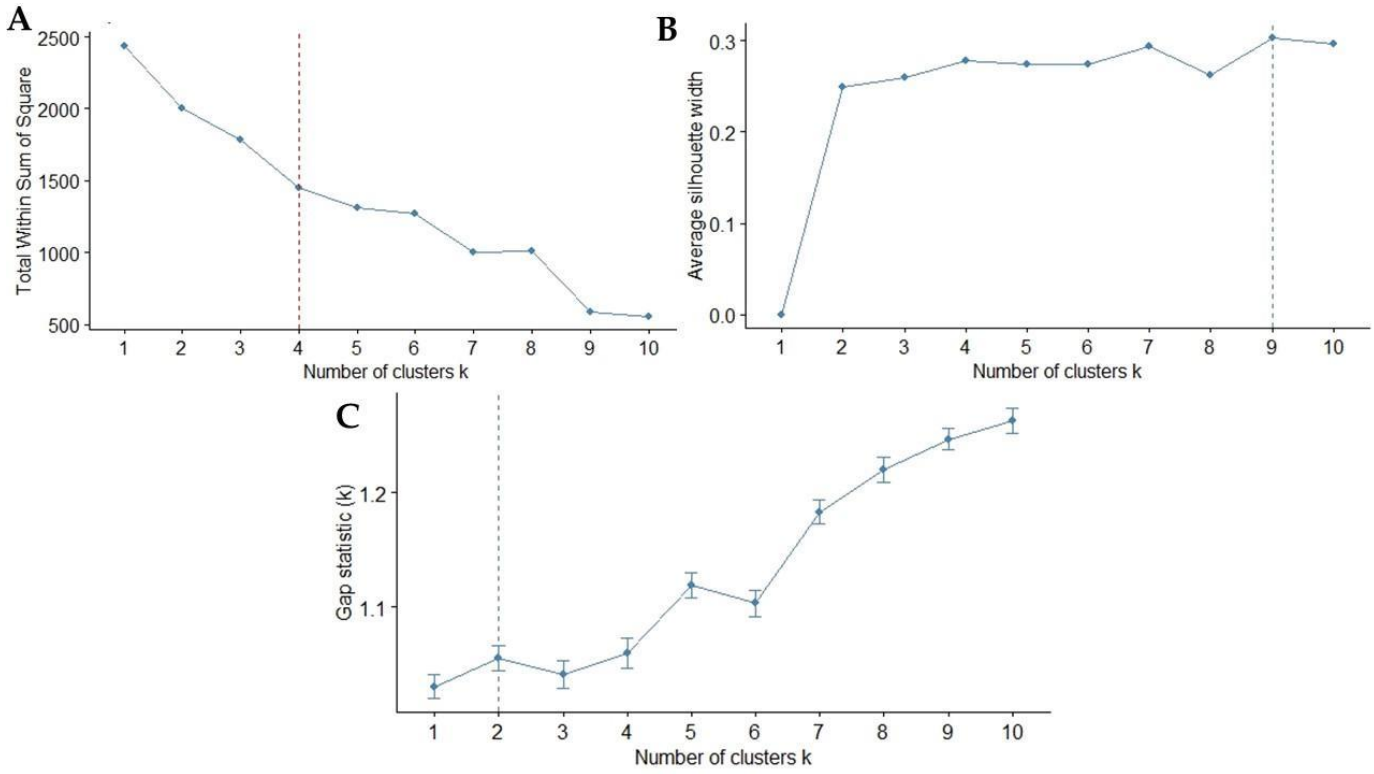

Figura 5. Cálculo de número óptimo de *clusters* (*k*) según el método **A)** del codo **B)** Silhouette y **C)** Gap Statistic.

Al hacer el análisis de clusterización se encontró que 2 *clusters* fueron los más óptimos para separar o caracterizar los pacientes según la información registrada (fig. 6). El cluster 1 agrupó 192 individuos, mientras que el segundo agrupó 215, siendo el primero el que presenta una mayor variabilidad, representada por líneas más alejadas del centroide en comparación con el cluster 2.

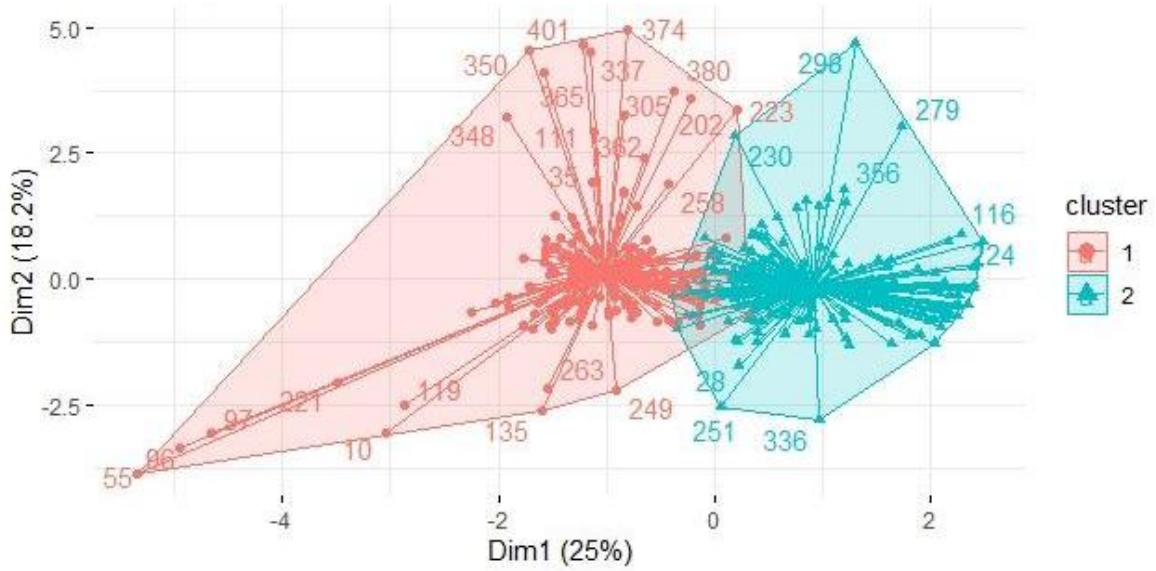

Figura 6. Clusterización de los datos con dos *k* especificados mediante el método *K-means*.

Los resultados mostraron que el cluster 2 agrupó a pacientes mucho más jóvenes con edades de entre 0 y 59 años, que no estaban vacunados. Adicionalmente, un menor número de reinfectados se observó en este cluster. Las diferencias observadas de estas variables fue significativa entre los dos clusters (**tablas 3 y 4**). Estos resultados sugieren que pacientes jóvenes sin ningún tipo de inmunidad, ya sea por vacunación o infección natural previa, siguen expuestos ante la aparición de nuevas variantes de SARS-CoV-2, pues no se observó diferencia entre la variante causante de la infección (*p=0.83*) en ambos grupos.

Por su parte, el *cluster* 1 agrupó a pacientes con un estado de vacunación adecuado y con edades mayores a 60 años. Curiosamente, aunque la mayoría de pacientes de este grupo presentaban 2 dosis de la vacuna Sinovac y el clúster 2 agrupó más casos de no vacunados, no se encontró una diferencia significativa entre el número de fallecidos en cada clúster (*p= 0.42*). Estos resultados muestran que dos dosis de la vacuna Sinovac no logran proteger de estados graves de la enfermedad o de muerte en los pacientes infectados después de aproximadamente 220 días de su aplicación.

No se encontró diferencia estadística de los síntomas y antecedentes clínicos entre los dos clusters (**tabla S8).**

Tabla 3. Caracteristicas epidemiológicas encontradas en cada uno de los clusters obtenidos mediante el método de K-means.

Variable	Hospitalizados				
	Clúster 1 n=192		Clúster 2 n=215		*p-value*
	n	%	n	%	
Edad (años), mediana (IQR)	75 (67 - 83)		56 (23 - 74)		**< 0.05**
Grupo de edad					
0-9	0	0,00	32	14,88	
10-17	0	0,00	6	2,79	
18-39	6	3,13	47	21,86	**< 0.05**
40-59	17	8,85	31	14,42	
> 60	169	88,02	99	46,05	
Procedencia					
Antioquia	35	18,23	39	18,14	
Atlántico	44	22,92	67	31,16	0.70
Valle del Cauca	113	58,85	109	50,70	
Género					
Femenino	85	44,27	99	46,05	
Masculino	107	55,73	116	53,95	0.99
Días de Hospitalización					
Mediana (IQR)	5 (3 - 8)		5 (1.5 - 5)		**< 0.05**
Variante					
Delta	32	16,67	46	21,40	
Omicron	160	83,33	169	78,60	0.83
Desenlace					
Vivo	158	82,29	183	85,12	
Fallecido	34	17,71	32	14,88	0.42
Infección previa					
No	176	91,67	207	96,28	
Si	16	8,33	8	3,72	**< 0.05**
Valor de *Ct*					
Mediana (IQR)	22.06 (19.21 - 24.96)		23.0 (20 - 25.19)		
Promedio (DE)	22.08		22.66		0.20

Tabla 4. Estado vacunal de los pacientes agrupados en cada uno de los clusters obtenidos mediante el método de *K-means*.

Variable	Hospitalizados TOTAL				
	Cluster 1 n=192		Cluster 2 n=215		*p-value*
	n	%	n	%	
Antecedente de Vacunación					
Si	191		56		
No	1		159		**< 0.05**
Estado de vacunación					
Optima	4		1		
Adecuada	171		39		**< 0.05**
No vacunado	17		175		
Dosis Aplicadas					
0	1		159		
1	25		22		**< 0.05**
2	162		33		
3	4		1		
Tipo de vacuna					
AstraZeneca	21		12		
Johnson & Johnson - Janssen	9		6		**< 0.05**
ARNm-1273 - Moderna	4		9		
BioNTech, Pfizer	44		13		
CoronaVac - Sinovac	113		16		
No vacunado	1		159		

Modelos de *Machine Learning*

ML fue utilizado para investigar los efectos cooperativos de las variables demográficas, vacunales y clínicas de los pacientes para determinar cuales pacientes podrían morir y por ende ayudar a priorizar su tratamiento mientras están hospitalizados. En nuestros datos, los pacientes fallecidos tuvieron un bajo porcentaje dentro del dataset total (16.25%); para evitar que los modelos fueran más exitosos prediciendo a pacientes vivos, se realizó un balanceo de datos mediante el método de sobremuestreo *Random Over-Sample* (ROS), en el cual se duplican las observaciones minoritarias para equilibrar los datos entre la variable respuesta. La base final obtenida tuvo 341 pacientes fallecidos y 341 vivos.

Los modelos empleados fueron Logistic regression, K-Nearest-Neighbor, Decision Tree, Random Forest, Support Vector Machine, Naïve Bayes, Support Gradient Boosting, Stochastic Gradient Descent y AdaBoost classifier; los hiperparametros usados fueron los predeterminados para cada modelo. En la figura 7 se muestran las curvas ROC obtenidas y la tabla 4 resume la precisión (*accuracy*) de cada uno de los modelos. Se obtuvo que el mejor modelo para predecir el estado final de los pacientes fue bosque aleatorio (*Ramdon Forest*), con una presición del 0.94, superando a los modelos de árbol de desiciones y *Gradient boosting*, los cuales tuvieron una presición del 0.88 y 0.83, respectivamente. Para validar la estabilidad de los modelos con un *accuracy* superior al 80% se realizó validación cruzada con 5 iteraciones. Se obtuvo que el mejor modelo sigue siendo bosque aleatorio, demostrando que no había *overfitting* del modelo (tabla 5).

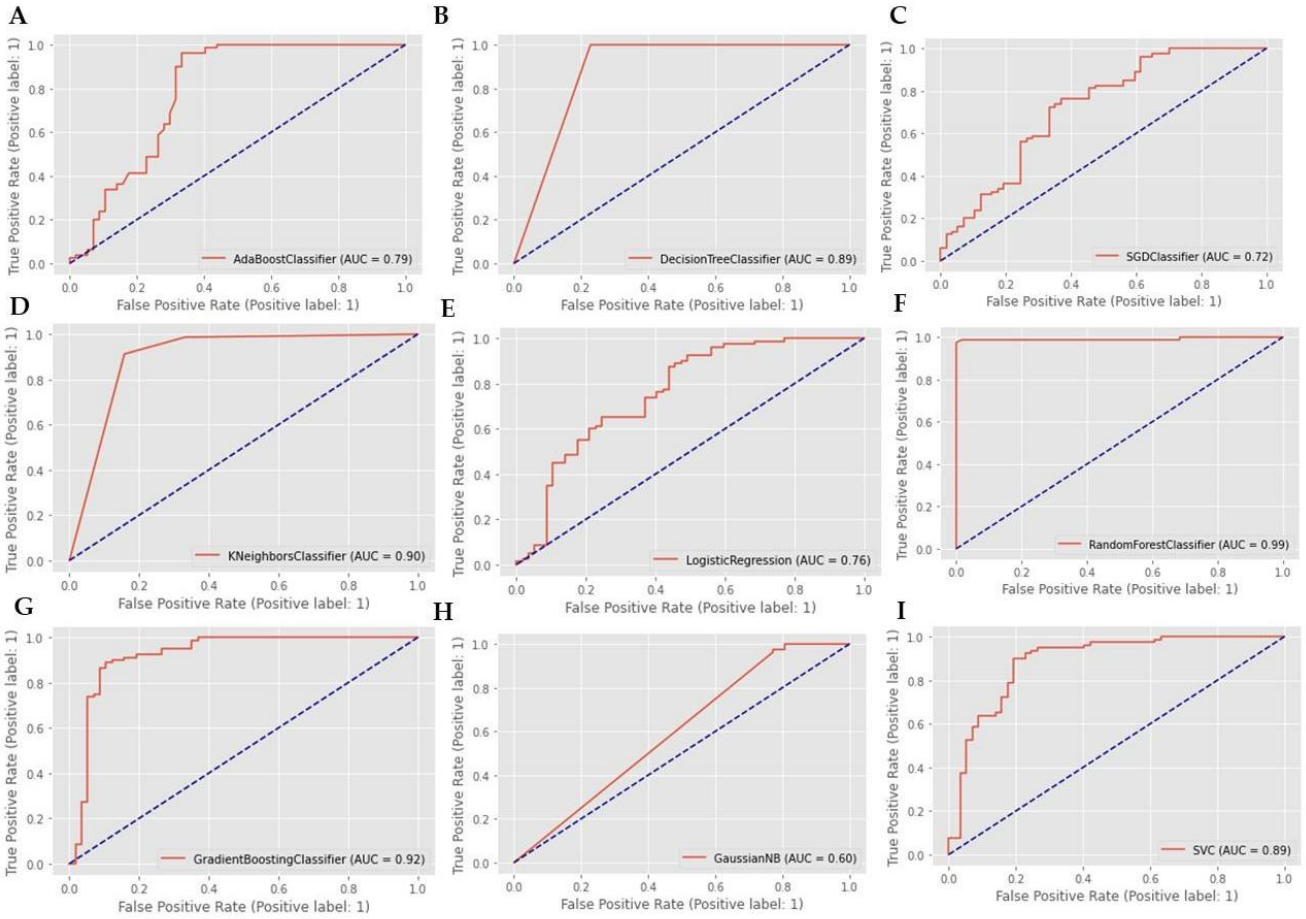

Figura 7. Curvas ROC de los modelos de machine learning para la predicción del desenlace de pacientes hospitalizados por COVID-19 obtenidas con los moelos **A)** AdaBoost classifier, **B)** Decision Tree, **C)** Support Gradient Boosting, **D)** K-Nearest-Neighbor, **E)** Logistic regression, **F)** Random Forest, **G)** Support Gradient Boosting, **H)** Naïve Bayes y **I)** Support Vector Machine.

Tabla 4. Resultados de sensibilidad, especificidad y precisión de los modelos de ML creados con la información analizada.

Modelo	Verdadero Negativo (TN)	Falso Positivo (FP)	Falso Negativo (FN)	Verdadero Positivo (TP)	Valor predictivo positivo (PPV) o Sensibilidad	Valor predictivo negativo (NPV) o Especificidad	*Accuracy*
Logistic Regression	36	21	22	58	73,4%	63,2%	72,26%
KNN Classification	42	15	7	73	83,0%	73,7%	85,40%
Decision Tree	44	13	3	77	85,6%	77,2%	90,51%
Random Forest	52	5	3	77	93,9%	91,2%	94,16%
Support Vector Machine	38	19	9	71	78,9%	66,7%	84,67%
Support Naive Bayes	11	46	2	78	62,9%	19,3%	65,69%
Support Gradient Boosting	40	17	5	75	81,5%	70,2%	85,40%
Stochastic Gradient Decent	28	29	7	73	71,6%	49,1%	67,88%
AdaBoost	38	19	11	69	78,4%	66,7%	74,45%

Tabla 5. Resultados de la validación cruzada de los mejores modelos obtenidos.

	Validación Cruzada 5-Pliegues					
Modelo	**Iteración 1**	**Iteración 2**	**Iteración 3**	**Iteración 4**	**Iteración 5**	**Media VC**
KNN Classification	0,83	0,74	0,83	0,81	0,84	0,81
Decision Tree	0,89	0,87	0,94	0,87	0,92	0,90
Random Forest	0,96	0,92	0,95	0,94	0,95	0,95
Support Vector Machine	0,74	0,78	0,83	0,83	0,87	0,81
Support Gradient Boosting	0,80	0,85	0,87	0,86	0,93	0,86

DISCUSIÓN

El machine learning, como herramienta de la inteligencia artificial, puede acelerar los procesos de detección y monitoreo de infecciones por SARS-CoV-2, asi como su prevención y tratamiento (Mottaqi et al., 2021). Además, el uso de estos análisis mejora sustancialmente los procesos de análisis y desarrollo de información relacionada en la etapa post-vacunación de la pandemia. En el ambito médico, el uso de ML no intenta reemplazar las tareas del personal médico, pero se ha demostrado que estas son tecnologías prometedoras que dan como resultado una mejor potencia de procesamiento, confiabilidad y que superan tareas específicas de atención médica, por lo que el desarrollo de nuevos modelos proveen una gran ayuda para la toma de desiciones en salud (Davenport & Kalakota, 2019; Lalmuanawma, Hussain, & Chhakchhuak, 2020; Yilmaz & Tolk, 2008). En este estudio implementamos técnicas de machine learning para identificar perfiles de los pacientes hospitalizados por COVID-19 durante la entrada de la variante Omicron a Colombia.

La mayoría de paciente con COVID-19 son sintomáticos y múltiples estudios han caracterizado los las caracteristicas clínicas de esta enfermedad. (Fu et al., 2020) realizaron un metaanálisis con distintos reportes y encontraron que los síntomas más comunes fueron fiebre, tos y fatiga, lo que concuerda con lo observado en este estudio. Aunque nuestros resultados mostraron diferencias significativas de las variables clínicas entre pacientes fallecidos y vivos, las proporciones de estas variables después de realizar la clusterización no presentó diferencias entre los grupos formados. La clusterización mediante el algoritmo de *K-means* es una técnica no probabilistica que permite identificar grupos o clusters de los datos en un espacio multidimensional (Bishop, 2006) y permite extraer interpretaciones que no son posibles por herramientas computacionales convencionales y concluir sobre la continua generación de datos de SARS-CoV-2 (Mottaqi et al., 2021). Debido a la gran diversidad de síntomas y a otros factores clínicos de las personas infectadas, es importante investigar el rol de estos en el desenlace de los pacientes con COVID-19. Dos clusters basados en la edad, valor de Ct del gen y en información de hospitalización y vacunación fueron obtenidos en este trabajo. (Molina-Mora et al., 2022) realizaron una clusterización basada en los síntomas y comorbilidades durante la etapa de pre-vacunación y encontraron que los aspectos clínicos del paciente y los genotipos de SARS-CoV-2 estuvieron presentes en todos los grupos, sin un patrón particular. Estos resultados muestran que los síntomas y/o antecedentes clínicos deben ser analizados con preacución a la hora de priorizar pacientes hospitalizados

con COVID-19, pues por si solos no permiten identificar pacientes con mayor riesgo de muerte en la etapa post-vacunación. Este tipo de información es útil para informar a los proveedores de salud y a los tomadores de desiciones en salud pública en su esfuerzo por controlar brotes de nuevas variantes de SARS-CoV-2 y dismminuir la mortalidad de los casos (Fu et al., 2020).

Cuando se compararon los pacientes estratificados por la variante causante de la infección, se encontró que la edad fue la única variable con una diferencia significativa, donde la variante Delta afectó a pacientes más jovenes en comparación con los casos de Omicron. Sin embargo, el uso de ML no supervisado mediante clusterización mostró que los clusters formados estuvieron igualmente afectados por ambas variantes. Debido a que tampoco hubo una diferencia estadística entre la proporción de fallecidos en cada cluster, nuestros resultados sugieren que la variante de SARS-CoV-2 por si sola no es un factor de riesgo para el fallecimiento en los pacientes por grupos de edad específicos. (Nakamichi et al., 2021) encontraron que la información la variante genética no mejora las predicciones durante la enfermedad, sugiriendo que esta contribuye de manera mínima en la determinación del desenlace final del paciente.

Distintos factores que influyen en la enfermedad pueden ser específicos en distintas poblaciones, por lo tanto estudios particulares son requeridos en cada región geográfica (Molina-Mora et al., 2022). A diferencia de las demás variables, en este trabajo se encontró que en la etapa post-vacunación de Colombia, una edad avanzada (>60 años) y falta de inmunidad previa (natural o inducida) hacen parte de los factores de riesgo en pacientes hospitalizados con COVID-19. Estudios previos han demostrado que hay una gran reducción en la protección contra la variante Omicron en comparación con la variante Delta tras la vacunación primaria, y que dosis de refuerzo aumentan la efectividad de protección contra estados severos de la enfermedad causadas por las variantes Omicron y Delta (Eggink et al., 2022; Thompson et al., 2022). Nuestros resultados mostraron un gran porcentaje de pacientes hospitalizados que tenían historial de vacunación con un tiempo mayor a 200 días sin dosis de refuerzo, esto se debe a que la efectividad de las vacunas se ve afectada con el tiempo debido a la disminución de la inmunidad inducida y/o a una evasión del sistema inmune de las nuevas variantes declaradas *VOC* (Thompson et al., 2022). Esto también explica el bajo porcentaje (4.9%) de pacientes con un estado de vacunación óptimo (3 dosis) entre los pacientes hospitalizados. Nuestros análisis y estudios *in vitro* realizados por nuestro grupo (Álvarez-Díaz et al., 2022) fueron fundamentales para promover la aplicación de la tercera dosis en pacientes mayores

tras la entrada de la variante Omicron al país y para seguir con las medidas de bioseguridad por un mayor tiempo como recomendó el Ministerio de Salud de Colombia durante el cuarto pico epidemiológico.

El enfoque de análisis de estos datos mediante ML permite modelar una mejor forma de predecir el desenlace de paceintes COVID-19. De manera similar, (Nakamichi et al., 2021) obtuvieron que, entre distintos modelos evaluados, *Random Forest* obtuvo resultados superiores para predecir la hospitalización en pacientes infectados. Asimismo, encontraron que la información epidemiológica más caracteristicas clínicas permitieron un valor de área bajo la curva de 0.93; la variante no mejoró el desempeño del modelo.

El alto poder de precisión para predecir el estado final de pacientes infectados con SARS-CoV-2 después del periodo de vacunación del modelo creado con *Ramdon Forest*, muestra que las variables incluidas para este análisis deben ser tenidas en cuenta a la hora de priorizar pacientes hospitalizados con COVID-19 y analizadas para la toma de desiciones en salud pública. Es importante seguir testeando estos modelos a medida que aparezcan nuevas variantes con el objetivo comprobar e identificar rápidamente patrones subyacentes ante infecciones causadas por nuevas variantes de este virus.

CONCLUSIÓN

El presente estudio pretendió caracterizar pacientes hospitalizados con COVID-19 durante la entrada de la variante Omicron en Colombia mediante el uso de machine learning como método de análisis rápido y profundo de los datos obtenidos. Se encontró que los principales factores de riesgo en la etapa post-vacunación en el país fueron edades mayores a 60 años y falta de inmunidad, así como falta de dosis de refuerzo. Aunque los métodos univariados asociaron a la variante Delta con pacientes más jovenes, los análisis multidimensionales mediante ML permitieron identificar que las variantes estuvieron igualmente distribuidas entre los grupos de pacientes identificados. Los pacientes que tenían anteccedente de vacunación tenian una mediana de tiempo de 220 días, sugiriendo que la efectividad de las vacunas disminuye después de este tiempo. Nuestros resultados fueron útiles para la toma de desiciones de salud pública del país, promoviendo dosis de refuerzo y el mantenimiento de otras medidas durante el cuarto pico epidemiológico de infecciones por SARS-CoV-2, producido por la variante de preocupación Omicron. Este estudio demuestra que las técnicas de machine learning pueden ser aplicadas exitosamente para el análisis de brotes producidos por nuevas variantes de SARS-CoV-2 con el objetivo de encontrar patrones característicos de la enfermedad producida con mayor rapidez y asi ayudar a una adecuada y rápida toma de desiciones y tratamiento de pacientes.

Bibliografía

Álvarez-Díaz, D. A., Muñoz, A. L., Herrera-Sepúlveda, M. T., Tavera-Rodríguez, P., Laiton-Donato, K., Franco-Muñoz, C., … Mercado-Reyes, M. (2022). Neutralizing responses in fully vaccinated with BNT162b2, CoronaVac, ChAdOx1, and Ad26.COV2.S against SARS-CoV-2 lineages in Colombia, 2020-2021. *MedRxiv*, 2022.03.15.22272371. https://doi.org/10.1101/2022.03.15.22272371

Amat, J. (2016). Machine learning con Python y Scikit-learn. Retrieved June 13, 2022, from https://www.cienciadedatos.net/documentos/py06_machine_learning_python_scikitlearn.html

Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., & Lloyd, S. (2017). Quantum machine learning. *Nature 2017 549:7671*, *549*(7671), 195–202. https://doi.org/10.1038/nature23474

Bishop, C. M. (2006). Pattern Recogniton and Machine Learning. In *Information Science and Statistics*. Retrieved from https://www.springer.com/gp/book/9780387310732

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthc J*, *6*(2), 94–98. https://doi.org/10.7861/FUTUREHOSP.6-2-94

Desy, P., Ernesto, A., Sahararini, A. F., Evandyano, G., Nasution, B. I., & Intan, J. (2022). *Implementation K-Means Clustering Based on Delta Variant and Omicron Variant by Sub-Districts in Jakarta*.

dos Santos, W. G. (2021). Co-infection, re-infection and genetic evolution of SARS-CoV-2: Implications for the COVID-19 pandemic control. *Journal of Cancer Biology*, *2*(3), 56–61. https://doi.org/10.46439/cancerbiology.2.025

Eggink, D., Andeweg, S. P., Vennema, H., van Maarseveen, N., Vermaas, K., Vlaemynck, B., … Knol, M. J. (2022). Increased risk of infection with SARS-CoV-2 Omicron BA.1 compared with Delta in vaccinated and previously infected individuals, the Netherlands, 22 November 2021 to 19 January 2022. *Eurosurveillance*, *27*(4). https://doi.org/10.2807/1560-7917.ES.2022.27.4.2101196

Fu, L., Wang, B., Yuan, T., Chen, X., Ao, Y., Fitzpatrick, T., … Zou, H. (2020). Clinical characteristics of coronavirus disease 2019 (COVID-19) in China: A systematic review and meta-analysis. *Journal of Infection*, *80*(6), 656–665. https://doi.org/10.1016/J.JINF.2020.03.041/ATTACHMENT/1D2BFA53-818B-

44D0-9772-8F1037FB181B/MMC1.DOC

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1), 100. https://doi.org/10.2307/2346830

Khattree, R., & Naik, D. N. (2002). Andrews plots for multivariate data: some new suggestions and applications. *Journal of Statistical Planning and Inference*, *100*(2), 411–425. https://doi.org/10.1016/S0378-3758(01)00150-1

Lalmuanawma, S., Hussain, J., & Chhakchhuak, L. (2020, October 1). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos, Solitons and Fractals*, Vol. 139, p. 110059. https://doi.org/10.1016/j.chaos.2020.110059

MathWorks. (2022). Curva ROC - MATLAB & Simulink. Retrieved June 13, 2022, from https://la.mathworks.com/discovery/roc-curve.html

Molina-Mora, J. A., González, A., Jiménez-Morgan, S., Cordero-Laurent, E., Brenes, H., Soto-Garita, C., … Duarte-Martínez, F. (2022). Clinical Profiles at the Time of Diagnosis of SARS-CoV-2 Infection in Costa Rica During the Pre-vaccination Period Using a Machine Learning Approach. *Phenomics*, (0123456789). https://doi.org/10.1007/s43657-022-00058-x

Mottaqi, M. S., Mohammadipanah, F., & Sajedi, H. (2021, January 1). Contribution of machine learning approaches in response to SARS-CoV-2 infection. *Informatics in Medicine Unlocked*, Vol. 23, p. 100526. https://doi.org/10.1016/j.imu.2021.100526

Nakamichi, K., Shen, J. Z., Lee, C. S., Lee, A., Roberts, E. A., Simonson, P. D., … Van Gelder, R. N. (2021). Hospitalization and mortality associated with SARS-CoV-2 viral clades in COVID-19. *Scientific Reports*, *11*(1). https://doi.org/10.1038/S41598-021-82850-9

OMS. (2022). Seguimiento de las Variantes. *2022*, 6. Retrieved from https://www.who.int/es/activities/tracking-SARS-CoV-2-variants

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Thompson, M. G., Natarajan, K., Irving, S. A., Rowley, E. A., Griggs, E. P., Gaglani, M., … Ong, T. C. (2022). Effectiveness of a Third Dose of mRNA Vaccines Against COVID-19-Associated Emergency Department and Urgent Care Encounters and Hospitalizations Among Adults During Periods of Delta and Omicron Variant Predominance - VISION Network, 10 States, August 2021-

January 2022. *MMWR. Morbidity and Mortality Weekly Report*, *71*(4), 139–145. https://doi.org/10.15585/MMWR.MM7104E3

Virgantari, F., & Faridhan, Y. E. (2020). K-Means Clustering of COVID-19 Cases in Indonesia's Provinces. *INTERNATIONAL JOURNAL OF ENGINEERING AND NATURAL SCIENCE*, *5*(2).

Wang, R., Chen, J., Gao, K., Hozumi, Y., Yin, C., & Wei, G.-W. (2020). Characterizing SARS-CoV-2 mutations in the United States. *Research Square*. https://doi.org/10.21203/RS.3.RS-49671/V1

Yilmaz, L., & Tolk, A. (2008). Intelligent Decision Making: An AI-Based Approach. *Intelligent Decision Making: An AI-Based Approach*, *97*, 193–226. https://doi.org/10.1007/978-3-540-76829-6

Zhou, H.-Y., Cheng, Y.-X., Xu, L., Li, J.-Y., Tao, C.-Y., Ji, C.-Y., … Wu, A. (2021). Genomic evidence for divergent co-infections of SARS-CoV-2 lineages. *BioRxiv*, 2021.09.03.458951. https://doi.org/10.1101/2021.09.03.458951

ANEXOS

Tabla S1. Porcentaje de valores faltantes de la base de datos total.

VARIABLES	n	%
Variante	838	64
Dias_refuerzo_infeccion	135	10
Nombre_biologico	131	10
Biológico_refuerzo	130	10
Dias_vacuna_1_Infección	114	9
Numero_Dosis	93	7
ESTADO_VACUNA	93	7
Vacunacion	92	7
Desenlace	58	4
Fiebre	53	4
Desnutricion	53	4
Odinofagia	53	4
Dif_Respiratoria	53	4
Adinamia	53	4
Asma	53	4
Cancer	53	4
EPOC	53	4
Diabetes	53	4
Tos	53	4
Fumador	53	4
Obesidad	53	4
Efermedad_cardiaca	53	4
Insuf_renal	53	4
Hipertension	53	4
Tuberculosis	53	4
Otros_dx	53	4
Cual_diag	53	4
Comorbilidades	53	4
VIH	53	4
Edad_cat	50	4
Edad	50	4
Departamento	0	0
Dias_diferencia_infecciones	0	0
Dias_hospitalizacion	0	0
CtGen_E	0	0
Sexo	0	0
Reinfectado	0	0

Tabla S2. Correlación entre las distintas variables numéricas presentes en el set de datos.

	Edad	CtGen_E	Dias_vacuna_1_Infección	Dias_refuerzo_infeccion	Dias_hospitalizacion	Dias_diferencia_infecciones
Edad	1,00	0,09	0,34	0,10	0,08	-0,01
CtGen_E	0,09	1,00	0,00	-0,06	-0,04	0,00
Dias_vacuna_1_Infección	0,34	0,00	1,00	0,15	0,03	0,06
Dias_refuerzo_infeccion	0,10	-0,06	0,15	1,00	0,02	-0,03
Dias_hospitalizacion	0,08	-0,04	0,03	0,02	1,00	-0,07
Dias_diferencia_infecciones	-0,01	0,00	0,06	-0,03	-0,07	1,00

Tabla S3. Datos epidemiológicos de pacientes infectados por las variantes Delta y Omicron.

Variable	**Hospitalizados**				
	Casos de Delta n=81		**Casos de Ómicron n=397**		*p-value*
	n	**%**	**n**	**%**	
Edad (años), mediana (IQR)	61 (36 - 76)		68 (50 - 82)		0,07
Grupo de edad					
0-9	6	7,41	34	8,56	
10-17	0	0,00	7	1,76	
18-39	20	24,69	37	9,32	
40-59	12	14,81	47	11,84	
> 60	43	53,09	262	65,99	
ND	0	0,00	10	2,52	**<0.05**
Procedencia					
Antioquia	21	25,93	71	17,88	
Atlántico	1	1,23	137	34,51	
Valle del Cauca	59	72,84	189	47,61	**<0.05**
Género					
Masculino	13	16,05	40	10,08	
Femenino	8	9,88	22	5,54	0,70
Días de Hospitalización					
Mediana (IQR)	6 (3 - 11)		3 (1-6)		
Estado Final					
Vivo	68	83,95	330	83,12	
Muerto	13	16,05	67	16,88	0,86

Infección previa					
No	75	92,59	375	94,46	
Si	6	7,41	22	5,54	0,52
Valor de *Ct*					
Mediana (IQR)	25.3 (20.9 - 30.1)		25.2 (20.6- 30.6)		0,93
Promedio (DE)	24.6 (8.1)		24.6 (8.3)		
ND	0		1		

Tabla S4. Antecedentes vacunales de pacientes infectados por las variantes Delta y Omicron. No se observó diferencia significativa entre ambas variantes.

Variable	**Hospitalizados TOTAL**				
	Casos de Delta n=81		**Casos de Ómicron n=397**		*p-value*
	n	**%**	**n**	**%**	
Antecedente de Vacunación					
Si	47	58,02	241	60,71	
No	33	40,74	132	33,25	0,22
ND	1	1,23	24	6,05	
Estado de vacunación					
Optima	2	2,47	20	5,04	
Adecuada	35	43,21	192	48,36	
No vacunado	43	53,09	160	40,30	0,07
ND	1	1,23	25	6,30	
Dosis Aplicadas					
0	33	89,19	132	62,26	
1	13	35,14	40	18,87	
2	32	86,49	180	84,91	
3	2	5,41	20	9,43	0,10
ND	1	2,70	25	11,79	
Tipo de vacuna					
AstraZeneca	5	13,51	31	12,86	
Johnson & Johnson - Janssen	3	8,11	12	4,98	
ARNm-1273 - Moderna	3	8,11	14	5,81	
BioNTech, Pfizer	14	37,84	45	18,67	
CoronaVac - Sinovac	21	56,76	125	51,87	0,23
ND	2	5,41	38	15,77	
No vacunado	33	89,19	132	54,77	

Tabla S5. Síntomas y comorbilidades estratificados entre los pacientes infectados por las variantes Delta y Omicron. No se observó diferencia significativa entre ambas variables.

Variable	Hospitalizados TOTAL				
	Casos de Delta n=81		Casos de Ómicron n=397		*p-value*
	n	%	n	%	
Tos					
Si	32	39,51	144	36,27	
No	48	59,26	242	60,96	
ND	1	1,23	11	2,77	0,65
Fiebre					
Si	43	53,09	190	47,86	
No	37	45,68	196	49,37	
ND	1	1,23	11	2,77	0,55
Odinofagia					
Si	66	81,48	319	80,35	
No	14	17,28	67	16,88	
ND	1	1,23	11	2,77	0,72
Dif_Respiratoria					
Si	53	65,43	233	58,69	
No	27	33,33	153	38,54	
ND	1	1,23	11	2,77	0,44
Adinamia					
Si	39	48,15	228	57,43	
No	41	50,62	158	39,80	
ND	1	1,23	11	2,77	0,17
Asma					
Si	78	96,30	380	95,72	
No	2	2,47	6	1,51	
ND	1	1,23	11	2,77	0,61
EPOC					
Si	72	88,89	348	87,66	
No	8	9,88	38	9,57	
ND	1	1,23	11	2,77	0,72
Diabetes					
Si	66	81,48	325	81,86	
No	14	17,28	61	15,37	
ND	1	1,23	11	2,77	0,67
VIH					
Si	79	97,53	383	96,47	
No	1	1,23	3	0,76	

ND	1	1,23	11	2,77	0,66
Efermedad_cardíaca					
Si	75	92,59	351	88,41	
No	5	6,17	35	8,82	
ND	1	1,23	11	2,77	0,52
Cáncer					
Si	75	92,59	368	92,70	
No	5	6,17	18	4,53	
ND	1	1,23	11	2,77	0,60
Desnutrición					
Si	79	97,53	381	95,97	
No	1	1,23	5	1,26	
ND	1	1,23	11	2,77	0,72
Obesidad					
Si	75	92,59	368	92,70	
No	5	6,17	18	4,53	
ND	1	1,23	11	2,77	0,60
Insuf_renal					
Si	78	96,30	362	91,18	
No	2	2,47	24	6,05	
ND	1	1,23	11	2,77	0,30
Hipertensión					
Si	60	74,07	271	68,26	
No	20	24,69	115	28,97	
ND	1	1,23	11	2,77	0,50
Tuberculosis					
Si	79	97,53	385	96,98	
No	1	1,23	1	0,25	
ND	1	1,23	11	2,77	0,34
Fumador					
Si	76	93,83	374	94,21	
No	4	4,94	12	3,02	
ND	1	1,23	11	2,77	0,50

Tabla S6. Datos epidemiológicos estratificados por pacientes vivos y fallecidos.

Variable	Hospitalizados				
	Fallecidos n=197		Vivos n=1062		*p-value*
	n	%	n	%	
Edad (años), mediana (IQR)	76 (67 - 86)		66 (43 - 78)		**<0.05**
Grupo de edad					
0-9	1	0,51	71	6,69	
10-17	0	0,00	21	1,98	
18-39	8	4,06	140	13,18	
40-59	17	8,63	176	16,57	
> 60	168	85,28	639	60,17	
Vacio	3	1,52	15	1,41	**<0.05**
Procedencia					
Antioquia	81	41,12	283	26,65	
Atlantico	43	21,83	196	18,46	
Valle del Cauca	73	37,06	583	54,90	**<0.05**
Género					
Femenino	81	41,12	468	44,07	
Masculino	116	58,88	594	55,93	0,44
Días de Hospitalización					
Mediana (IQR)	6 (3 - 11)		4 (2-8)		**<0.05**
Variante					
Mu	0	0,00	1	0,09	
Delta	13	6,60	68	6,40	
Omicron	67	34,01	330	31,07	
Linaje no asignado	117	59,39	663	62,43	0,83
Infección previa					
No	192	97,46	986	92,84	
Si	5	2,54	76	7,16	**0,02**
Valor de *Ct*					
Mediana (IQR)	25 (21 - 30)		26 (22- 30)		0,9799
Promedio (DE)	25 (6)		26 (6)		
ND	0		1		

Tabla S7. Antecedentes vacunales estratificados por pacientes vivos y fallecidos.

Variable	Hospitalizados TOTAL				
	Fallecidos n=197		Vivos n=1062		*p-value*
	n	%	n	%	
Antecedente de Vacunación					
Si	72	36,55	356	33,52	
No	118	59,90	627	59,04	**<0.05**
ND	7	3,55	79	7,44	
Estado de vacunación					
Optima	13	6,60	50	4,71	
Adecuada	92	46,70	491	46,23	
No vacunado	85	43,15	441	41,53	0,16
ND	7	3,55	80	7,53	
Dosis Aplicadas					
0	72	68,57	356	65,80	
1	20	19,05	119	22,00	
2	85	80,95	457	84,47	
3	13	12,38	50	9,24	0,23
ND	7	6,67	80	14,79	
Tipo de vacuna					
AstraZeneca	22	20,95	94	26,40	
Johnson & Johnson - Janssen	7	6,67	34	9,55	
ARNm-1273 - Moderna	5	4,76	31	8,71	
BioNTech, Pfizer	15	14,29	148	41,57	
CoronaVac - Sinovac	63	60,00	289	81,18	0,11
ND	13	12,38	110	30,90	
No vacunado	72	68,57	356	100,00	

Tabla S8. Síntomas y antecedentes clínicos estratificados por cada uno de los clusters obtenidos mediante el método de K-means.

Variable	Hospitalizados TOTAL				
	Fallecidos n=197		Vivos n=1062		*p-value*
	n	%	n	%	
Tos					
Si	66	33,50	367	34,56	
No	131	66,50	661	62,24	
ND	0	0,00	34	3,20	0,03
Fiebre					
Si	99	50,25	571	53,77	
No	98	49,75	457	43,03	
ND	0	0,00	34	3,20	0,02
Odinofagia					
Si	156	79,19	845	79,57	
No	41	20,81	183	17,23	
ND	0	0,00	34	3,20	0,02
Dif_Respiratoria					
Si	97	49,24	649	61,11	
No	100	50,76	379	35,69	
ND	0	0,00	34	3,20	0,00
Adinamia					
Si	92	46,70	587	55,27	
No	105	53,30	441	41,53	
ND	0	0,00	34	3,20	0,00
Asma					
Si	194	98,48	1011	95,20	
No	3	1,52	17	1,60	
ND	0	0,00	34	3,20	0,03
EPOC					
Si	171	86,80	923	86,91	0,03
No	26	13,20	105	9,89	
ND	15	7,61	46	4,33	
Fumador					
Si	186	94,42	992	93,41	
No	11	5,58	36	3,39	
ND	0	0,00	34	3,20	0,01