Facultad de Ingeniería

Escuela de Ingeniería Civil en Bioinformática

Inteligencia artificial para evaluar asociación de parámetros sistémicos y clínicos a los diferentes grados de severidad y nivel de secuela de pacientes que cursaron COVID-19

Kevin Patricio Aguilar Valdés

Tutora: PhD. Estefanía Nova Lamperti

Co-tutora: Mabel Vidal Miranda

Profesor Informante: PhD. José Reyes Suarez

Diciembre, 2021

CONSTANCIA

La Dirección del Sistema de Bibliotecas a través de su unidad de procesos técnicos certifica que el autor del siguiente trabajo de titulación ha firmado su autorización para la reproducción en forma total o parcial e ilimitada del mismo.

Talca, 2022

1. Contenido

2. ÍNDICE DE TABLAS

3. ÍNDICE DE ILUSTRACIONES

4. RESUMEN

La enfermedad de coronavirus 2019 (COVID-19) es una enfermedad infecciosa causada por el coronavirus del síndrome respiratorio agudo severo 2 (SARS-CoV-2), la cual se ha extendido a 192 países, incluido el nuestro, logrando 268.484.455 contagios confirmados y un total de 5.286.786 muertes a nivel mundial. Esta enfermedad se ha caracterizado por presentar diversos síntomas de acuerdo con la gravedad del cuadro infeccioso, pudiendo éste presentarse de forma asintomática, leve, moderada o severa, llegando incluso a causar un desenlace fatal. Además, se ha demostrado que los pacientes que logran sobrevivir presentan diversas secuelas que persisten meses posterior al término de la etapa aguda de la enfermedad. Actualmente, no está del todo clara la relación entre las características que presenta un paciente diagnosticado con COVID-19 con el grado de severidad que presentará durante la infección del virus, así como las posibles secuelas que pueda desarrollar una vez haya cursado la enfermedad.

El objetivo de este trabajo fue analizar diferentes parámetros clínicos, experimentales y demográficos para determinar su asociación a los distintos grados se severidad y nivel de secuela de pacientes que cursaron COVID-19. La metodología para abordar este proyecto fue a partir de información obtenida de 60 pacientes con diagnóstico confirmado de infección por SARS-CoV-2 que cursaron la patología con sintomatología leve (n=18), moderada (n=17) y severa (n=25). En primera instancia se configuró el conjunto de datos a partir de la información obtenida de cada participante. A continuación, y mediante métodos de selección de características, se determinó aquellos parámetros que tienen mayor relevancia al momento de asociarlos a los distintos grados de severidad y nivel de secuela. Finalmente, se realizó la evaluación de diferentes modelos supervisados de inteligencia artificial que permitieron clasificar grado de severidad y nivel de secuela en pacientes que cursaron COVID-19. A partir de este trabajo se obtuvieron parámetros que tienen mayor influencia en el desarrollo de niveles de severidad y secuela como lo son información clínica, de Tomografía Axial Computarizada (TAC), cuestionarios, espirometría y citoquinas plasmáticas. Además, se lograron generar modelos de clasificación con un desempeño de hasta 100% a partir de los métodos

de Random Forest y XGBoost, logrando aportar al análisis de información generado por la pandemia de COVID-19. En conclusión, métodos de inteligencia artificial permitieron identificar patrones y las características más relevantes que están asociadas al grado de severidad y nivel de secuela en pacientes que cursaron COVID-19.

5. ABSTRACT

The coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2). This virus has spread to 192 countries, including ours, and so far, 268,484,455 infections and 5,286,786 deaths worldwide have been confirmed. Asymptomatic infection or mild, moderate, severe or critical symptoms during infection have been reported in COVID-19 patients. In addition, it has been shown that patients who survive exhibit diverse sequelae, which persist for months after the acute stage of the disease. Currently, the relationship between COVID-19 patient characteristics and the degree of severity and sequelae during and after viral infection is not entirely clear. Thus, the aim of this study was to analyze different clinical, experimental and demographic parameters to determine their association at different levels of severity and sequelae in COVID-19 patients. Clinical, experimental, and demographic data obtained from 60 patients with a confirmed diagnosis of mild (n = 18), moderate (n = 17) and severe (n = 25) SARS-CoV-2 infection was obtained. The dataset was configured from the information obtained from each participant. Subsequently, and through methods of characteristics' selection, those parameters that are most relevant when associating them with the different levels of severity and sequelae were determined. Then, the evaluation of different supervised artificial intelligence models was carried out that allow classifying the degree of severity and level of sequelae in patients who had COVID-19. Our results revealed that the most important features in the modeling of the degree of severity and sequelae were clinical information, Computerized Axial Tomography (CT), questionnaires, spirometry, and plasma cytokines. In addition, classification models with a performance of up to 100% was generated using the Random Forest and XGBoost methods, contributing to the analysis of information generated by the COVID-19 pandemic. In summary, artificial intelligence methods made it possible to identify patterns and the most relevant characteristics that are associated with the degree of severity and level of sequelae in patients with COVID-19.

6. INTRODUCCIÓN

El síndrome respiratorio agudo severo conocido como (SARS) por sus siglas en inglés (severe acute respiratory syndrome coronavirus) es un cuadro causado por el coronavirus (SARS-CoV) que comenzó a ser estudiado como una neumonía atípica en la provincia de Guangdong, China, en el año 2002 [1]. Sin embargo, estudios genéticos posteriores demostraron su naturaleza vírica y su relación con la familia *Coronaviridae*. Este virus, transmisible de persona a persona, se propagó a 11 países, alcanzando un total de 8.448 infectados y 774 muertos en la que fue considerada la primera pandemia del siglo XXI [2,3].

Posteriormente, y al igual que su antecesor, el SARS-CoV-2 tuvo su origen en China, más precisamente en Wuhan, Provincia de Hubei, en diciembre de 2019 [4]. Estudios preliminares sugerían un origen zoonótico, apuntando a algunas especies de murciélagos como potenciales transmisores de virus relacionados al SARS-CoV (SARS-CoVs) a los humanos. Esta problemática desencadenó en la necesidad de caracterizar este nuevo patógeno, logrando establecer un 76,6% de identidad de secuencia entre SARS-CoV y SARS-CoV2, presentando este último un 96,2% de identidad con coronavirus de murciélago [5]. De esta forma, lo que nuevamente comenzó como el estudio de una neumonía de etiología no definida, ya en enero de 2020 fue declarada por la Organización Mundial de la Salud (OMS) como una epidemia de emergencia sanitaria de preocupación internacional. En febrero del mismo año la OMS nombra esta enfermedad 2019-nCoV como enfermedad de coronavirus 19 (COVID-19), que a 64 días de su primer contagiado ya contaba con 77.401 casos confirmados de infecciones en China, sobrepasando el brote de SARS-CoV en 2002 [6]. Actualmente, la ya declarada pandemia alcanza 268.484.455 contagios y suma un total de 5.286.786 muertes a nivel mundial [7].

6.1. Características y transmisión del SARS-CoV2

En cuanto a las vías de transmisión del SARS-CoV2, estas incluyen fómites y partículas que pueden transmitirse mediante contacto estrecho y sin protección entre portadores del virus, que pueden ser sintomáticos o asintomáticos, e

individuos sanos [8]. Estudios recientes indagan las dinámicas de transmisión de COVID-19 y establecen un periodo medio de incubación de 5 a 6 días y hasta 19 días en casos particulares. Sin embargo, las autoridades internacionales determinaron un periodo de cuarentena de 14 días para los individuos infectados [9].

La manifestación más característica de la enfermedad de COVID-19 son fiebre y tos seca, además de otros signos más graves dependiendo del grado de severidad y del desarrollo de la enfermedad en el paciente, expuestas en la Tabla 1[10,11]. Generalmente, los pacientes presentan síntomas similares a la neumonía, siendo los adultos mayores de sexo masculino la población más propensa a contraer el virus[12].

Tabla 1. *Manifestaciones clínicas en tres niveles diferentes de severidad en COVID-19*

Manifestación clínica	**Fiebre, tos seca, dificultad para respirar, dolor muscular, confusión, dolor de cabeza, dolor de garganta, rinorrea, dolor de pecho, diarrea, náuseas, vómitos, escalofríos, esputo, hemoptisis, disnea, neumonía bilateral, anorexia, dolor de pecho, leucopenia, linfopenia, trastornos olfatorios y del gusto, niveles altos de citocinas plasmáticas en pacientes UCI (IL2, IL7, IL10, GSCF, IP10, MCP1, MIP1A y TNFa)**	
Diferentes niveles de COVID-19	**Leve**	Fiebre, tos, opacidad de vidrio esmerilado (GGO), no-neumonía y neumonía leve.
	Moderado	Disnea grave, saturación de oxígeno en sangre ≤93%, frecuencia respiratoria ≥30/min, presión parcial de oxígeno arterial a fracción de oxígeno inspirado <300 y/o infiltrados pulmonares >50% en 24 a 48 horas. Necesario UCI.
	Severo	Síndrome de dificultad respiratoria aguda (SDRA), insuficiencia respiratoria, choque séptico y/o disfunción o insuficiencia de múltiples órganos, acidosis metabólica, disfunción de coagulación.

Dado al aumento significativo de los casos a nivel mundial se ha hecho necesario establecer protocolos de testeo para lograr identificar, aislar y tratar a los individuos que son portadores del virus. Esto con el fin de reducir las tasas de mortalidad y el riesgo de aumentar los casos positivos. Existe una serie de métodos entre los que

destacan la reacción en cadena de polimerasa con transcriptasa inversa (RT-PCR), la RT-PCR en tiempo real (rRT-PCR) y la amplificación isotérmica mediada por bucle de transcripción reversa (RT-LAMP). Estos ensayos poseen una sensibilidad similar al momento de detectar la presencia del virus en el individuo, siendo una pieza fundamental para combatir el avance de la pandemia [13–15].

6.2. Factores de riesgo al cursar COVID-19

Se ha observado en el estado del arte que enfermedades preexistentes en pacientes contagiados actúan como factores de riesgo que pueden ser claves en la severidad del cuadro infeccioso. Padecimientos como insuficiencia cardiaca crónica, hipertensión, epilepsia, diabetes y obesidad son algunas de las principales afecciones asociadas al riesgo de muerte en pacientes de COVID-19 [16,17]. Esta información a sido utilizada para definir estrategias que prioricen la vacunación de determinados grupos acorde a las posibilidades de desarrollar un cuadro severo de la enfermedad[18].

6.2.1. Obesidad

La obesidad, por sí sola considerada una epidemia [19], ha jugado un rol negativo en la actual pandemia. Las diferencias anatómicas, fisiológicas y fenotípicas de los pacientes con obesidad los hacen más propensos a desarrollar un cuadro severo de infección [20]. Por otro lado, estudios recientes exponen como la función respiratoria se compromete en pacientes obesos que cursan COVID-19, produciendo incluso una manifestación prolongada de la enfermedad, aumentando su riesgo de muerte [21].

6.2.2. Diabetes

Siendo una enfermedad que está asociada a la anterior, la diabetes es otro factor de riesgo que contribuye a la severidad y mortalidad de pacientes con COVID-19 [22]. En la literatura se observa evidencia de que existe una alta susceptibilidad de algunas enfermedades infecciosas de origen bacteriana en pacientes con diabetes. Esto ocurre debido a una respuesta inmune desregulada [23], provocando que un

porcentaje importante de pacientes diabéticos sean hospitalizados por la infección del virus.

6.2.3. Afecciones cardiacas

Cada una de la amplia gama de afecciones cardiacas incrementa por si sola los riesgos de sufrir complicaciones al estar contagiado de COVID-19 [24]. Numerosos estudios han demostrado complicaciones sufridas a causa de la infección del patógeno, así como cifras elevadas de muertes por falla cardiaca [25–27].

6.2.4. Enfermedades mentales

Al igual que enfermedades que se manifiestan físicamente, las enfermedades mentales se incluyen en los padecimientos de riesgo en pacientes de COVID-19. La población mayor, especialmente aquellos con demencia senil, son un grupo vulnerable y en riesgo de contraer la enfermedad, presentar síntomas graves y malos resultados de evolución, mostrando tasas elevadas de infección y letalidad[28]. Además de esto, estudios exponen las complicaciones sufridas por pacientes que padecen de Parkinson y el cómo han tenido que ajustar sus terapias causando cambios farmacocinéticos en su organismo, produciendo complicaciones y afecciones severas en su aparato digestivo [29].

6.3. Minería de Datos e Inteligencia artificial

El almacenamiento de datos se ha vuelto una tarea rutinaria de los sistemas informáticos en el amplio abanico de intereses. Los datos obtenidos son de especial importancia para quienes los manejan ya que muchas veces contienen información valiosa que actúa como una memoria de su trabajo. Memoria que puede ser accedida para obtener información de ella, siendo este el objetivo de la minería de datos [30]. En la toma de decisiones durante la pandemia, el análisis de datos ha sido fundamental desde diferentes perspectivas, tanto en salud como en economía, y por esta razón es esencial contar con sistemas robustos y automatizados.

Data mining es definido como el proceso iterativo de extracción de patrones predictivos ocultos en enormes bases de datos utilizando tecnologías de inteligencia

artificial, aprendizaje automático y técnicas de estadística [31]. Han sido innumerables disciplinas e industrias las que se han visto beneficiadas por el uso de la información para obtener datos de mercado, producción, atención al cliente, entre otras (Tabla 2). Todas impulsadas a través de inteligencia artificial, definida por Kaplan y *cols*. como la capacidad de un sistema que logra incorporar datos externos, aprender de ellos y utilizar lo aprendido para lograr tareas específicas a través de un proceso de adaptación flexible [32]. Esta tecnología se ha visto potenciada debido al progreso en la obtención de datos digitalizados, aprendizaje automático y desarrollo de mejor infraestructura informática [33]. Logrando así posicionarse como una herramienta útil en el análisis de datos a gran escala y participando en la automatización de procesos productivos y de control de fabricación [34].

Tabla 2. *Resumen áreas de aplicación Inteligencia Artificial [35]*

Área de aplicación	***Subcampo***	***Métricas de evaluación***	***Estudio***
Minería de texto	Clasificación de texto	Exactitud Valor-F Precisión Exhaustividad	(Forman, 2003)[36]
	Clúster de texto	Entropía Precisión	(Liu *et. Al*, 2003)[37]
Procesamiento de imagen/ visión por computadora	Clasificación de imagen	Error cuadrático Medio	(Bins *et al*., 2001)[38]
	Clasificación de densidad	Exactitud	(Muštra *et al*., 2017)[39]
Bioinformática	Descubrimiento de biomarcadores	Estabilidad Área bajo la curva ROC (AUC)	(Dessi *et al*., 2013)[40]
	Clasificación de expresión génica en microarreglos	Exactitud	(Abusamra, 2013)[41]
Industrial	Diagnóstico de fallas	Exactitud	(Liu *et al*., 2014)[42]

6.3.1. Aprendizaje Automático (Machine Learning)

Machine Learning es la disciplina científica enfocada en como las computadoras aprenden de los datos [43,44]. Este enfoque busca identificar relaciones o patrones en los datos a través de algoritmos de cómputo eficientes [45]. Los principales tipos de

aprendizajes utilizados por las computadoras se encuentran subclasificados en aprendizaje supervisado y aprendizaje no supervisado.

Aprendizaje supervisado: como su nombre lo indica, este tipo de aprendizaje implica algoritmos que aprenden bajo la presencia de un supervisor. Es decir, al entrenar una máquina o computadora, el aprendizaje supervisado se refiere a una categoría de métodos en los que enseñamos o entrenamos un algoritmo utilizando datos, de manera de guiar el algoritmo con etiquetas asociadas a los datos [44]. Los problemas recurrentes en los que se utiliza este método incluyen reconocimiento de escritura, clasificación de imágenes o documentos. Comúnmente se utiliza en problemas de clasificación, eligiendo entre los subgrupos que mejor describan la instancia del dato y su predicción [46].

Aprendizaje no supervisado: este tipo de aprendizaje no cuenta con una salida definida. El método intenta identificar los patrones de ocurrencia a través del comportamiento de los datos sin información o conocimiento a priori. Sin embargo, a pesar de que no se cuenta con información o etiquetas de los datos, se cuenta con la información en sí. Esto significa que se puede extraer referencias de las observaciones de los datos que utilicen como entrada. Es decir, se intenta encontrar estructuras y patrones significativos en las observaciones para extraer características generativas y con fines exploratorios [44].

6.3.2. Selección de características

Existe una gran variedad de artículos en la literatura que exponen diferentes estrategias de como se ha aplicado inteligencia artificial para el estudio de patologías crónicas [47,48], estudios clínicos [49], diseño de drogas [50,51], entre otros. Estos estudios buscan realizar predicciones entrenando modelos con datos obtenidos en las distintas áreas, utilizando características de las que solo algunas pueden estar relacionadas con el concepto de estudio. En este escenario, la selección de características es una parte importante para acelerar el aprendizaje y generar una reducción de información, así como para mejorar la claridad del concepto [52].

Existe muchos algoritmos que incluyen la clasificación de variables como mecanismo de selección principal y auxiliar. Esto debido a su simplicidad, escalabilidad y nivel de éxito [53]. Se estipula un criterio de clasificación adecuado para puntuar las variables, posterior a eso se establece un umbral para descartar las características por debajo de esta marca, filtrando así las variables con mayor relevancia [54].

6.3.3. Modelos de predicción

La minería de datos puede abordar cualquier problema que contenga datos almacenados. En general, no existe una forma única para solucionar todos los problemas que se puedan plantear, este debe ser abordado de manera de aproximarse lo más que pueda a una posible solución. Existen una serie de técnicas para modelar las diferentes problemáticas, las más utilizadas son expuestas en la Tabla 3.

Tabla 3. *Breve descripción de técnicas de predicción* [55]

Técnica	Descripción
Clustering o agrupamiento	Utilizan medidas de proximidad entre individuos y a partir de ahí buscan grupos de individuos semejantes entre sí, según una serie de variables.
Redes Bayesianas	Se representan todos los posibles sucesos mediante un grafo de posibilidades condicionales de transición entre sucesos. Este permite establecer relaciones causales y efectuar predicciones.
Redes Neuronales	Son generalización de modelos estadísticos clásicos. Su potencial radica en el aprendizaje secuencial, el que utiliza transformaciones de variables originales de predicción y la no linealidad del modelo. Permite aprender sin precisar de formulación de un modelo concreto.
Árboles de decisión	Permiten representar de forma visual las reglas de decisión por las cuales operan los individuos, a partir de datos históricos. Su principal ventaja es la facilidad de interpretación.
Series de tiempo	A partir de una serie de comportamientos históricos, se permite modelar las componentes básicas de la serie, su tendencia, ciclo y estacionalidad. Esto para poder predecir por ejemplo la cifra de una venta o el consumo de un producto o servicio.

6.4. Aplicación de Inteligencia Artificial en estudios de COVID-19

Analizar desde diferentes perspectivas toda la información que se pueda obtener desde los pacientes que han cursado COVID-19 ha sido clave. De esta manera, se han descubierto patrones relevantes que han permitido establecer indicadores esenciales para guiar la toma de decisiones relativas a la pandemia y su posterior tratamiento [56].

Como una manera de procesar complejos conjuntos de datos, la inteligencia artificial se ha introducido como una gran herramienta que incluso nos permite realizar predicciones de comportamientos en diferentes situaciones [57], automatizando tareas que requieren la misma inteligencia que si fuesen hechas por seres humanos a través del aprendizaje automático y aprendizaje profundo [58].

El aprendizaje automático ha sido una gran contribución durante la pandemia, como por ejemplo en estudios que evalúan la utilización de modelos de predicción, empleando datos clínicos y de laboratorio, para mejorar el rendimiento en que la prueba RT-PCR y la tomografía computarizada (TC) determinan a los huéspedes del patógeno [59]. Gangloff *y cols*., entrenaron tres modelos de predicción: regresión logística, random forest, y redes neuronales. Gracias a estos modelos de predicción, se logró incrementar las prestaciones diagnósticas ya sea en TC de tórax, con una exactitud de 95%, y RT-PCR, con un 92.4% de exactitud, para el diagnóstico de COVID-19.

Otra aplicación es un estudio realizado por Banerjee y *cols*. en el Hospital Israelita Albert Einstein, en São Paulo donde implementaron modelos de aprendizaje automático aplicados a hemogramas completos suministrados por el Hospital [60]. Lograron predecir con un 95% de precisión patrones de respuesta del sistema inmune y cambios en los diferentes parámetros que permite medir el hemograma en un rápido análisis de sangre.

En otro enfoque, que apunta a la forma en que se propaga el virus, Hass *y cols*. realizaron un estudio que utiliza información geográfica para así explorar patrones espacio-temporales que pudiesen estar contribuyendo a acelerar la cantidad de

contagios [61]. En este trabajo se monitoreó a nivel de continente en Europa y nivel país en Dinamarca. Los autores utilizaron métodos estadísticos clásicos y de aprendizaje automático con el fin de explorar la relación entre los factores que podrían influir en la propagación del virus en la población. Estas variables están asociadas a la densidad de la población, niveles de contaminación del aire y patrones de conducta, que describen cada una un factor importante a considerar para realizar predicciones. El análisis determinó que las zonas con mayor densidad poblacional están más propensas a generar un aumento en los contagios, así como el funcionamiento y concurrencia de clientes a locales de reunión social. Finalmente se evaluaron diferentes modelos de predicción y se estimó que futuras investigaciones requieren incorporar más información respecto a las medidas adoptadas por la autoridad como el cierre de la región y la contaminación, siendo estas de carácter importante a la hora de determinar el comportamiento de la pandemia. Adicionalmente se plantea lo difícil que es predecir el comportamiento humano y la forma en que este cumple con las normas sanitarias que buscan ralentizar la propagación del virus, en este caso viéndose perjudicadas por la variable cooperación en la población.

6.5. Desafío planteado

La creciente cantidad de datos obtenidos del COVID-19 a través de los centros de salud de todo el mundo podrían seguir métodos avanzados de aprendizaje automático y profundo para analizar efectos terapéuticos en futuros pacientes. De esta manera se podría brindar una mejor atención e incluso predecir su hospitalización, contribuyendo al funcionamiento del recinto asistencial, así como a otras áreas de interés para el combate a la pandemia (Ilustración 1).

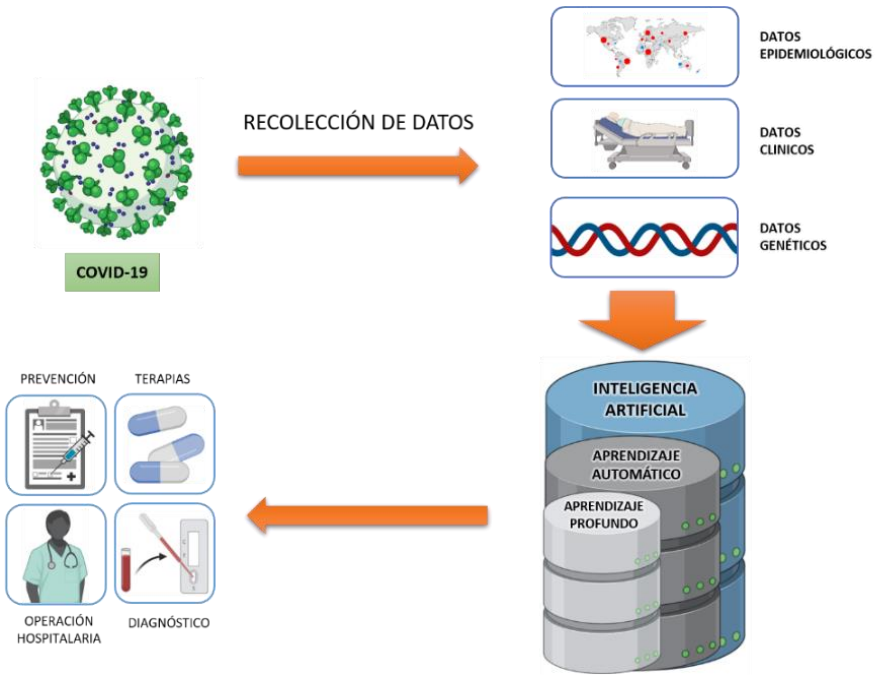

Ilustración 1. *Aplicación de inteligencia artificial y aprendizaje automático en la lucha contra el COVID-19.*

Una de las principales problemáticas que mantiene preocupados a los servicios de salud que combaten la pandemia, es la necesidad de identificar con precisión la severidad del cuadro infeccioso que tendrá el paciente dada la necesidad de brindarle un tratamiento acorde a sus requerimientos. Por otro lado, el aumento en el número de casos va de la mano con el uso de las unidades de tratamiento intensivo, infraestructura clave para atender y asistir a aquellos pacientes que desarrollan un cuadro severo. Es preciso determinar de forma oportuna la información clave en el progreso de la enfermedad, ensayos clínicos y experimentales a lo largo de la infección. Además, se han definido variadas secuelas posterior al cuadro agudo del COVID-19, pero no se tiene claro cuáles son las variables que contribuyen con este cuadro de sintomatología persistente, incluso 12 meses post infección.

Por lo tanto, en este trabajo se evaluaron datos clínicos, experimentales y demográficos para determinar la asociación entre los diferentes niveles de severidad en pacientes que cursaron COVID-19 y las secuelas generadas post

infección. Para esto, se utilizaron parámetros sistémicos obtenidos de 60 pacientes distribuidos en los tres grados de severidad (leve, moderado, severo) y 13 individuos control. El estudio pretende responder, ¿Cuáles son las principales características que influyen en el grado de severidad y nivel de secuela de un paciente con COVID-19?, y ¿Es posible generar un modelo que clasifique grado de severidad y nivel de secuela de un paciente con COVID-19? Bajo estas interrogantes se desarrolló este estudio intentando aportar al análisis de la información generado por la pandemia.

7. HIPÓTESIS

El análisis de datos mediante diferentes modelos de aprendizaje automático permite determinar patrones y características asociadas al grado de severidad y nivel de secuela en pacientes que cursaron COVID-19.

8. OBJETIVOS

8.1. Objetivo General

Analizar diferentes parámetros clínicos, experimentales y demográficos para determinar su asociación a los distintos grados de severidad y nivel de secuela en pacientes que cursaron COVID-19.

8.2. Objetivos Específicos

- Configurar un set de datos a partir de información clínica, experimental y demográfica obtenida desde 60 pacientes que cursaron COVID-19 e individuos control.
- Determinar las características más relevantes asociadas con los diferentes niveles de severidad y secuela de pacientes que cursaron COVID-19.
- Evaluar diferentes modelos de Inteligencia Artificial para clasificar grado de severidad y nivel de secuela en pacientes que cursaron COVID-19.

9. MATERIALES

9.1. Diseño del Estudio

Los datos utilizados para este trabajo provienen de un estudio observacional y de cohorte prospectivo denominado COVID1005. El protocolo del estudio fue previamente registrado en el registro ISRCTN (ID: ISRCTN16865246) y fue aprobado por el comité ético científico (CEC) del Servicio de Salud Bio (código: CEC113) (Anexo 1). Los pacientes fueron reclutados desde el Hospital Regional Dr. Guillermo Grant Benavente de Concepción y el Complejo Asistencial Dr. Víctor Ríos Ruiz de Los Ángeles siguiendo las directrices sugeridas por el tratado STROBE. Al momento de la inclusión en el estudio, se obtuvo el consentimiento informado firmado, y todos los métodos se realizaron de acuerdo con la declaración de Helsinki y las buenas prácticas clínicas.

Grupos de Estudio

Se incluyeron 60 pacientes con edad ≥ 18 años, con diagnóstico confirmado de infección por SARS-CoV-2 mediante PCR durante los meses de abril a julio de 2020. Los pacientes debían contar con registro clínico y seguimiento durante la fase aguda de la enfermedad por COVID-19. Se incluyeron pacientes con COVID-19 de severidad *leve* (n=18, sintomáticos, sin diagnóstico de neumonía o hipoxemia), *moderada* (n=17, sintomáticos, con diagnóstico de neumonía que requirió hospitalización, sin conexión a ventilación mecánica invasiva (VMI)) y *severa* (n=25, hipoxemia severa, con necesidad de VMI y estadía en Unidad de cuidados intensivos (UCI)). Se excluyeron aquellos pacientes con comorbilidad respiratoria previa (asma, enfermedad pulmonar obstructiva crónica, otra), pacientes mayores de 70 años, además de participantes con pérdida de seguimiento, historia de trasladado a otro hospital o ciudad posterior al alta, aquellos en cuidados paliativos o que tuvieran una discapacidad mental que impidiera completar las evaluaciones. Además, se incluyeron 13 potenciales individuos control que no presentaron COVID-19 confirmado por PCR negativo semanal y ausencia de anticuerpos anti-SARS-CoV2 al momento de toma de muestra.

Obtención de datos en fase aguda

Se extrajeron datos sobre demografía (edad, sexo, años de escolaridad, zona rural), antropometría (índice de masa corporal (IMC) (Kg/M^2)), hábitos sociales (consumo de tabaco y alcohol) y comorbilidades (hipertensión, resistencia a la insulina, diabetes mellitus tipo 2, hipotiroidismo, arritmia, enfermedad coronaria o accidente cerebrovascular) y consumo de medicamentos. También se extrajeron los peores parámetros de laboratorio más alterados durante el periodo agudo (ferritina (mg/dl), proteína C reactiva (PCR) (mg/dl), recuento de leucocitos ($x10^9$/ l), recuento de linfocitos ($x10^9$/l), dímero D (mg/dl), fibrinógeno (mg/dL), PaO2/FIO2) y las intervenciones realizadas durante la estadía en pacientes hospitalizados (CNAF, posición en decúbito prono vigil, uso de esteroides (dexametasona), anti-interleucina 6 (tocilizumab), antibióticos, uso de VMI, días en VMI, bloqueo neuromuscular (BNM), posición en decúbito prono, traqueostomía, días en la UCI y días totales en el hospital. Para el grupo de COVID-19 leve, se extrajeron los datos de los participantes con al menos una visita médica y parámetros de laboratorio iniciales después del diagnóstico. Este grupo recibió seguimiento clínico por teléfono y atención de apoyo.

Obtención de datos en fase de secuela a corto plazo

Durante las 12 y 24 semanas posteriores a la infección, todos los participantes fueron citados para una evaluación clínica que incluyó síntomas relacionados con el COVID-19, (Fiebre, cefalea, tos, anosmia, ageusia, mialgia, dolor abdominal y toráxico, diarrea, tos, odinofagia y polipnea). De manera adicional a los síntomas del periodo agudo, se evaluaron los siguientes síntomas de manera dirigida: Sensación de cansancio, miedo a adquirir nueva infección por SARS-CoV-2, la disminución de lívido, presencia de alopecia, aparición de parestesias y/o paresia de extremidades inferiores o superiores. La disnea fue medida por el *modified Medical Research Council* (mMRC) modificado. Se consultó por fatiga muscular mediante auto reporte y la versión binaria del cuestionario de fatiga de Chalder. Se consideró un valor de corte ≥ 4 puntos como fatiga severa. Durante la evaluación, los participantes completaron los siguientes cuestionarios: Escala de ansiedad y

depresión hospitalaria (HADS) y el inventario de depresión de Beck. Además, se aplicó el cuestionario *short form* (SF-12) para valorar la calidad de vida relacionada con la salud se presentan los resultados en el dominio de salud física y salud mental. Finalmente, se exploró el cambio en la calidad de vida relacionada con la salud (CVRS) antes y después de la infección por COVID-19. Para este objetivo, se utilizó una escala visual analógica con un rango de 0% (peor salud imaginable) a 100% (mejor salud imaginable). Un cambio ≥10% fue sugerente de un cambio significativo en la CVRS, similar a la literatura previa. Para evaluación de función pulmonar se midieron gases en sangre arterial (pH, paO2, pCO2, HCO3-, exceso de base, gradiente A-a), espirometría basal y 15 minutos después de la inhalación de 400 µg de salbutamol utilizando un espirómetro de flujo (CPF-S / D; Medical Graphics Inc., EE. UU.), la capacidad pulmonar de difusión de monóxido de carbono (DLCO) mediante un pletismógrafo (Elite PlatinumDL; Medical Graphics Inc, EE. UU.), una Tomografía Computarizada de tórax de alta resolución, usando un escáner multidetector de 16, 30 y 32 cortes (SOMATOM, Siemens, Alemania) y estudio de sueño y ciclo circadiano mediante astigrafía. Finalmente, parámetros inflamatorios y análisis de parámetros sanguíneos fueron medidos en los pacientes y en los individuos control mediante análisis de hemograma, perfil bioquímico completo, insulina, HOMA (Modelo de Evaluación de Homeostasis) , perfil hepático y perfiles inmunológicos como activación celular y presencia de anticuerpos específicos para proteínas del SARS-COV-2 en circulación.

9.2. Conjunto de datos

Para llevar a cabo los objetivos planteados en el presente escrito, se trabajará con archivos de datos en documentos Excel (*.xlsx*) otorgados por PROYECTO COVID1005 de acuerdo con lo anteriormente descrito. Estos archivos almacenan datos asociados a cada paciente que cursó COVID-19.

9.3. Herramientas de análisis de datos

- Sistema operativo Windows 10
- Lenguajes de programación: Python versión 3.9
- Software y paquetes adicionales:
  - Excel (https://www.microsoft.com/es-ww/microsoft-365/excel) Herramienta de cálculo, gráficas, tablas y lenguaje de programación implementados como hojas de caculo aplicadas al manejo de datos numéricos y representación de esquemas de datos.
  - Jupyter Notebook (https://jupyter-notebook.readthedocs.io/en/stable/) Entorno informático interactivo y servidor web. Posibilita la programación en varios idiomas. Presenta una lista ordenada de celdas de entrada o salida que pueden contener código, texto, gráficos o formulas. Permite la ejecución de celdas específicas, logrando el testeo rápido de código, así como el entorno de trabajo necesario para la actividad a realizar.

10. MÉTODOS

En esta sección se describe la metodología para lograr los objetivos planteados. Esta sigue esencialmente una serie de etapas, asociadas a cada objetivo específico, y procesos para cada una de estas, descritas en la Ilustración 2.

10.1. Metodología objetivo específico 1: Configurar un set de datos a partir de información clínica, experimental y demográfica obtenida desde 60 pacientes que cursaron COVID-19 e individuos control.

10.1.1. Preparación de datos

Los datos fueron trabajados mediante archivos Excel, consolidando el registro de los ensayos médicos e información demográfica antes mencionados. Se realizó un pre-procesamiento de los datos, identificando información faltante que fue evaluada para procesos de deleción o imputación de éstos. Además, se analizaron las unidades de medidas de la información obtenida de los datos clínicos, experimentales y demográficos, para realizar un proceso de estandarización de los registros, con la finalidad de obtener datos calculados bajo la misma medida y estos fueran expresados de forma cuantitativa.

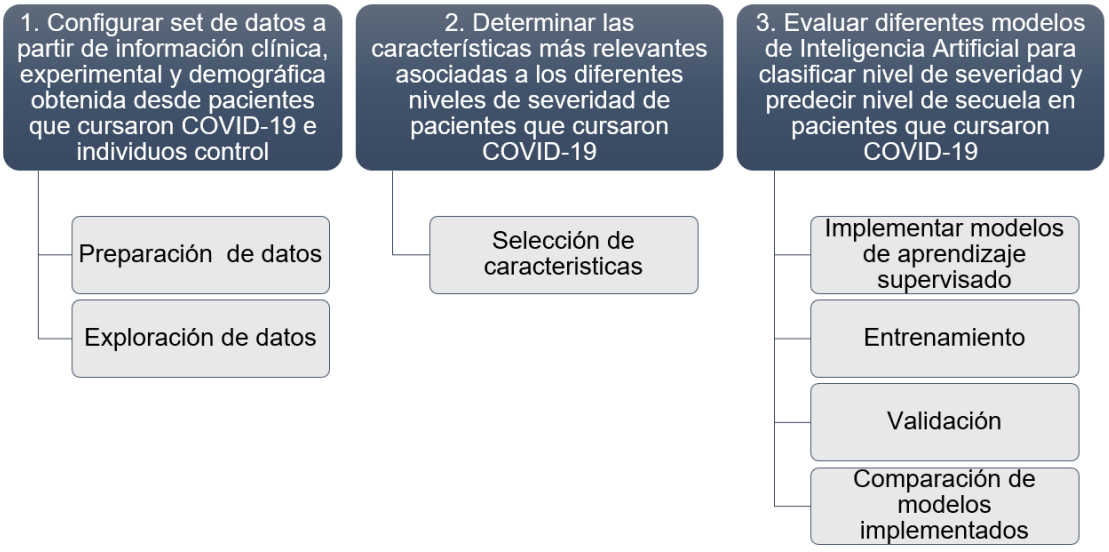

Ilustración 2. *Diagrama de flujo de la metodología utilizada.*

10.1.2. Exploración de datos

A partir del preprocesamiento de los datos fue posible generar un nuevo dataframe (estructura de datos con dos dimensiones) conteniendo toda la información para ser trabajada con las librerías *pandas* y *NumPy* en el entorno de programación *Jupyter Notebook*, a través del lenguaje de programación *Python*[62,63]. La librería *pandas* permitió manipular y analizar estructuralmente la información que se debió analizar. La librería *NumPy* añadió soporte para datos de alta dimensionalidad tanto en arreglos como matrices, operando de una manera eficiente con una gran colección de funciones matemáticas y estadísticas.

Los archivos proporcionados correspondían a datos ordenados que contenían información de muestras de serología, activación celular y función pulmonar, así como información demográfica obtenida de 60 pacientes infectados con COVID-19 y 13 pacientes control (que no han cursado COVID-19)[1].

En estos archivos las filas describen el identificador de cada paciente y las columnas el identificador de la muestra, expuestos en la Tabla I y Tabla II respectivamente en la sección 2 del Anexo del documento.

Se posee un número conocido de pacientes en cada nivel de severidad de la enfermedad, descritos en la Tabla 4. Las etiquetas del grado de severidad fueron indicadas por un equipo experto que clasificó previamente a los pacientes.

Tabla 4. *Grado de severidad de los pacientes que padecieron la enfermedad COVID-19*

Cohorte de pacientes COVID-19	Cantidad de pacientes
Severo	25
Moderado	17
Leve	18
Control	13

[1] Las muestras realizadas a los individuos de control difieren en cantidad con el resto de los pacientes.

Además, con las librerías *numpy* y *pandas* se realizó un análisis exploratorio a los datos de manera de visualizar y descubrir a priori ideas generales del comportamiento de los datos. Esto permitió identificar áreas o patrones en los que es importante profundizar en las siguientes etapas.

En este punto, también se evaluó la información obtenida a partir de los individuos sanos, logrando identificar que tan diferentes son respecto al resto de los pacientes.

10.2. Metodología objetivo específico 2: Determinar las características más relevantes asociadas a los diferentes niveles de severidad de pacientes que cursaron COVID-19.

10.2.1. Selección de características

Este proceso se define como la búsqueda y selección de las variables más útiles dentro de un *dataset*. Estas incluyen desde funciones que extraen características debido a que presentan cierta cantidad de valores nulos, hasta el uso de modelos matemáticos para determinar la relación entre las características y su potencial importancia dentro del set de datos. En este sentido, se utilizaron tres métodos que permitieron evaluar las características más relevantes asociadas a los diferentes grados de severidad y nivel de secuela en pacientes que cursaron COVID-19.

Feature Selection

Esta técnica asigna puntajes (*scores)* a las variables de entrada de un modelo predictivo. Este valor indica la importancia relativa de cada característica con las variables de destino, permitiendo identificar aquellas que aportan más información al modelo planteado. Para realizar esto, se utilizó la biblioteca de programación en Python *Scikit-learn*[64]. Esta implementa funciones que emplean técnicas de puntuación, que a su vez permite evaluar las características del set de datos con diferentes algoritmos definidos como de regresión, clasificación y permutación[65].

Mutual information

Esta técnica es parte de teoría de la información y es aplicada para obtener información acerca de la dependencia entre variables. *Mutual Information* es

calculada mediante dos variables, midiendo la reducción de la incertidumbre de una, dado un valor conocido de la otra, siendo equivalente a:

$$MI(X;Y) = H(X) - H(X|Y)$$

Donde la entropía (*H*) mide el nivel de incertidumbre esperado en una variable aleatoria. Por tanto, *H(X)* representa cuanta información se puede obtener de la variable aleatoria *X* al observar la variable *Y* [66]*.*

En el caso de este estudio se utilizó este procedimiento ya que permite generar *subsets* de características, identificando aquellas que son irrelevantes en proceso de obtener la variable destino. Para esto se implementó la clase *mutual_info_classif* de la biblioteca *Scikit-learn.*

ANOVA (Analysis of Variance)

Corresponde a un método estadístico de selección de características. Este evalúa la varianza entre dos o más grupos de información determinando el nivel de diferencia que presentan uno del otro. Este análisis asume la hipótesis:

H_0: la varianza de todos los grupos es igual.

H_1: al menos una varianza de los grupos es diferente.

Este método permitió evaluar la información entregada por cada característica y su posible similitud entre los registros de cada una, logrando discriminar y puntuar de acuerdo con las diferencias que presentan estos grupos de información. Para esto se implementó la clase *f_classif* de la biblioteca *Scikit-learn*.

SHapley Additive exPlanation

Este algoritmo emplea ingeniería inversa en los métodos de inteligencia artificial, logrando determinar el cómo se construye el modelo a partir de la información entregada por cada característica dentro del set de datos [67]. La metodología que

emplea es la de asignar valores considerando la importancia de cada característica por cada posible combinación de estas al momento de generar el modelo, entregando información de la importancia de cada característica por cada clase en el modelo. La implementación de este algoritmo permitió evaluar cuales son las características que más influyen en cada clase al momento de generar los modelos de clasificación. Este análisis fue realizado a través de la clase *SHAP*.

10.3. Metodología objetivo específico 3: Evaluar diferentes modelos de Inteligencia Artificial para clasificar grado de severidad y nivel de secuela en pacientes que cursaron COVID-19.

Un algoritmo de *Aprendizaje Automático* es un conjunto de cálculos que permiten generar un modelo a partir de un set de datos. Para obtener este modelo específico se realiza el análisis de los datos, logrando identificar patrones o tendencias en estos.

En este trabajo se implementaron algoritmos de clasificación para determinar la asociación de los parámetros (sistémicos y clínicos) en los grados de severidad y nivel de secuela de pacientes que cursaron COVID-19.

En esta sección se utilizaron los datos procesados en las tareas anteriores para generar los modelos, entrenarlos y validar la información obtenida de estos.

10.3.1. Modelos de aprendizaje automático a implementar.

Modelos de clasificación

De manera de aumentar el conocimiento estructural de los datos y determinar la asociación de los parámetros con el grado de severidad y nivel de secuela en los pacientes que cursaron COVID-19, se implementaron modelos de clasificación. Para el caso de este estudio se utilizaron cinco métodos de aprendizaje supervisado: Árboles de decisión [68], Random Forest[69], XGBoost[70], Naive bayes[71] y Support vector machines[72], descritos a continuación:

- *Arboles de decisión*: Es un método de aprendizaje automático de regresión y clasificación. Este construye un modelo de decisión basado en categorizar una

serie de condiciones que ocurren de manera sucesiva a partir de los valores obtenidos desde las características de los datos. Este modelo fue implementado a través de la clase *DecisionTreeClassifier* del módulo *scikit-learn*, capaz de ejecutar clasificaciones multiclase en un set de datos.

- *Random Forest:* Este método corresponde a los llamados *Ensemble Methods,* capaces de combinar predicciones a través de estimadores construidos con algoritmos de aprendizaje. Este método logra reducir la variación combinando diversos árboles de decisión. Para la predicción de un nuevo caso este es dirigido desde el nodo superior hasta el nodo terminal donde se le asigna una etiqueta. Este proceso es iterado por los demás árboles en el ensamblado, estableciendo como predicción aquella etiqueta que obtenga la mayor cantidad de incidencias. Este modelo fue implementado a través de la clase *RandomForestClassifier* del módulo *scikit-learn.*

- *XGboost:* Este método también corresponde a los llamados *Ensemble Methods.* Su funcionamiento es similar al de *Random Forest* y se basa en agregar árboles de decisión secuencialmente a fin de aprender desde los resultados obtenidos y corregir el error que estos producen hasta que dicho error no se pueda mejorar. Se ejecuta a través de un procesamiento en paralelo, realizando poda de árboles, manejo de datos perdidos y regularización para minimizar el sobreajuste o sesgo del modelo. Este modelo fue implementado a través de la clase *XGBoost*.

- *Naive Bayes:* Este método corresponde a un conjunto de algoritmos basados en la aplicación del *Teorema de Bayes*[73] con un supuesto de independencia entre los predictores, asumiendo que una característica en particular de una clase no está relacionada con la presencia de ninguna otra característica. Este algoritmo es reconocido por su función de predicción de múltiples clases, fácil implementación y rapidez para obtener resultados. Este modelo fue implementado a través de la clase *naive_bayes* del módulo *scikit-learn.*

- *Support Vector Machines (SVM):* este método es utilizado para clasificación y regresión. Este algoritmo busca definir un hiperplano capaz de separar en forma óptima los puntos de una clase y otra, que posiblemente han sido proyectados a un espacio dimensional mayor. Este modelo fue implementado a través de la clase *svm* del módulo *scikit-learn.*

Optimización de hiperparámetros

Los hiperparámetros corresponden a los parámetros ajustables que controlan el proceso de creación de un modelo de inteligencia artificial. Estos permiten ajustar el rendimiento de los métodos de clasificación y el cómo se construye el modelo a partir de los datos de entrenamiento. Este proceso se llevó a cabo con el fin de obtener la mejor combinación de parámetros por cada método utilizado en un intento por generar un modelo más preciso en cada caso. Para esto se utilizaron dos algoritmos de búsqueda de hiperparametros: *GridSearchCV* y *RandomizedSearchCV*, descritos a continuación:

- *GridSearch Cross Validation:* Este algoritmo genera una grilla de combinaciones entre los distintos valores que pueden tomar los parámetros de cada método de clasificación. Posteriormente entrena y evalúa iterativamente los modelos generados, determinando así cual combinación de parámetros generó la mejor estimación. Este proceso fue implementado mediante la clase *GridSearchCV* del módulo *sckit-learn*.

- *RandomizedSearch Cross Validation:* Este algoritmo genera valores aleatorios para los parámetros de cada método de clasificación, permitiendo evaluar valores intermedios como posibles estimadores del modelo generado. Fue implementado mediante la clase *RandomizedSearchCV* del módulo *scikit-learn*.

10.3.2. Entrenamiento de algoritmos de aprendizaje automático supervisado.

Los modelos de regresión y clasificación fueron entrenados a través del método de LOOCV (Leave One-Out Cross Validation)[74], el cual se utilizó para testear los resultados de un análisis estadístico y así garantizar que la subdivisión de los datos de entrenamiento y de prueba sean independientes. Este método de validación cruzada deja una muestra de los datos fuera, utilizando el resto para entrenar el modelo y posteriormente utilizar esta muestra para medir el rendimiento del algoritmo implementado. El procedimiento descrito se repite por cada uno de los ejemplos que contenga el set de entrenamiento, para finalmente integrar los resultados de todas estas evaluaciones. Esta sección de la metodología se implementó mediante el módulo en *Python model_selection.leaveoneout*.

10.3.3. Validación de algoritmos de aprendizaje supervisado.

Una vez realizado el entrenamiento de los algoritmos de aprendizaje automático estos fueron evaluados de acuerdo con sus estimaciones. Este proceso se realizó mediante la utilización de métricas de evaluación en función del proceso anteriormente propuesto, las cuales son:

Métricas de clasificación

Al ejecutar algoritmos de clasificación estos pueden entregar 4 tipos de resultados.

- *Verdaderos positivos (TP):* ocurren cuando se predice una observación en una clase y realmente pertenece a dicha clase.
- *Verdadero negativo (TN):* son cuando se predice que una observación no pertenece a una clase y efectivamente no pertenece a dicha clase.
- *Falso positivo (FP):* ocurren cuando se predice que una observación corresponde a una clase, pero realmente no pertenece a dicha clase.
- *Falso negativo (FN)*: es cuando se predice que una observación no pertenece a una clase y realmente pertenece a dicha clase.

Es a partir de estos resultados que se obtuvieron las cuatro principales métricas para evaluar los modelos de clasificación[75]:

- *Accuracy:* Se utilizó para medir el porcentaje de predicciones correctas para los datos de prueba. Se calculó dividiendo el número de predicciones correctas por el número de predicciones totales.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- *Precision: está dada por la fracción de observaciones TP entre los demás ejemplos que se predijo pertenecían a una determinada clase.*

$$Precision = \frac{TP}{TP + FP}$$

- *Recall:* es definida como la fracción de observaciones que fueron predichas para una clase con respecto a todos los ejemplos que realmente si pertenecían a esa clase.

$$Recall = \frac{TP}{TP + FN}$$

- *F1 Score:* se utiliza combinando *Precision* y *Recall* en un solo valor. Este cálculo permite comparar el rendimiento combinado de ambas métricas.

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

11. RESULTADOS

11.1. Resultados objetivo específico 1: Configurar un set de datos a partir de información clínica, experimental y demográfica obtenida desde 60 pacientes que cursaron COVID-19.

A partir del set de datos previamente descrito se realizó a modo exploratorio un análisis de registros para así determinar metodologías a seguir. En este punto se optó por descartar la información recabada de los 13 pacientes control. Esto debido a la diferencia en etiquetas con información que presentaban en comparación con el resto de los pacientes evaluados.

Paralelamente se revisaron aquellas características con información faltante, optando por su deleción en la mayoría de los casos, dejando así un set de datos completo de 60 registros y 327 características.

11.1.1. Sets de datos

Se determinó el trabajar además con el set de datos completo y con porciones de este, por lo que se generaron 10 *subsets* correspondientes a distintos grupos de características agrupados de acuerdo con los exámenes, ensayos y cuestionarios por los que fueron obtenidos, descritos por cantidad en la Tabla 5 y por grupo de características en la sección 3 del ANEXO.

Tabla 5. *Subsets de datos y su dimensión. Donde* ***n*** *representa la cantidad de pacientes y* ***x*** *la cantidad de características.*

Set de datos	Dimensión[n,x]
COVID-1005.xls	[73, 335]
FullSet	[60, 327]
TAC	[60, 16]
Clínica	[60, 25]
Hemograma	[60, 14]
Poligrafía	[60, 12]
Cuestionario	[60, 14]
Espirometría	[60, 33]
Actigrafía	[60, 21]
Demografía	[60, 35]
Sintomatología	[60, 14]
Sintomatología2	[60, 14]

11.1.2. Clase objetivo

Se asignaron tres clases para grado de severidad, correspondientes a la gravedad de los síntomas presentados por los pacientes de COVID-19. De la misma forma se establecieron 4 clases para determinar el nivel de secuela de los individuos estudiados, ambos expuestos en la Tabla 6.

Tabla 6. *Clases objetivos*

Clase objetivo	Dimensión	Clase
Severidad	[60,]	SEVERO
		MODERADO
		LEVE
Secuela	[60,]	TAC + DLCO
		TAC
		DLCO
		NO SEQ

TAC: Tomografía Axial Computarizada
DLCO: Capacidad de difusión pulmonar

11.2. Resultados objetivo específico 2: Determinar las características más relevantes asociadas a los diferentes grados de severidad y nivel de secuela en pacientes que cursaron COVID-19.

A partir de los subsets de datos obtenidos en el objetivo anterior es que se evaluó la importancia de características por cada grupo de información. Se utilizaron tres métodos de selección de características, evaluados para las dos clases objetivo de grado de severidad y nivel de secuela, obteniendo los resultados expuestos a continuación:

11.2.1. Set completo

Al evaluar la importancia de características del set completo en función de modelar el grado de severidad se observa que información clínica, de Tomografía Axial Computarizada (TAC) y cuestionarios son los datos que aportan mayor información a la generación del modelo de clasificación (Ilustración 3a). De la misma forma, al tener como objetivo el nivel se secuela, se observa un puntaje (score) más alto en aquellas características obtenidas a partir de exámenes de Tomografía Axial Computarizada (TAC), espirometría y citoquinas plasmáticas (Ilustración 3b).

Adicionalmente, en este punto se evaluó la importancia de características, para modelar nivel de secuela, del set completo excluyendo variables propias de exámenes TAC y Espirometría. Esto debido a que es a partir de estos dos sets es que se obtuvo la información para definir el nivel de secuela. Este análisis dio como

resultado que variables correspondientes a datos demográficos, antropométricos, cuestionarios y citoquinas plasmáticas (Ilustración 3c).

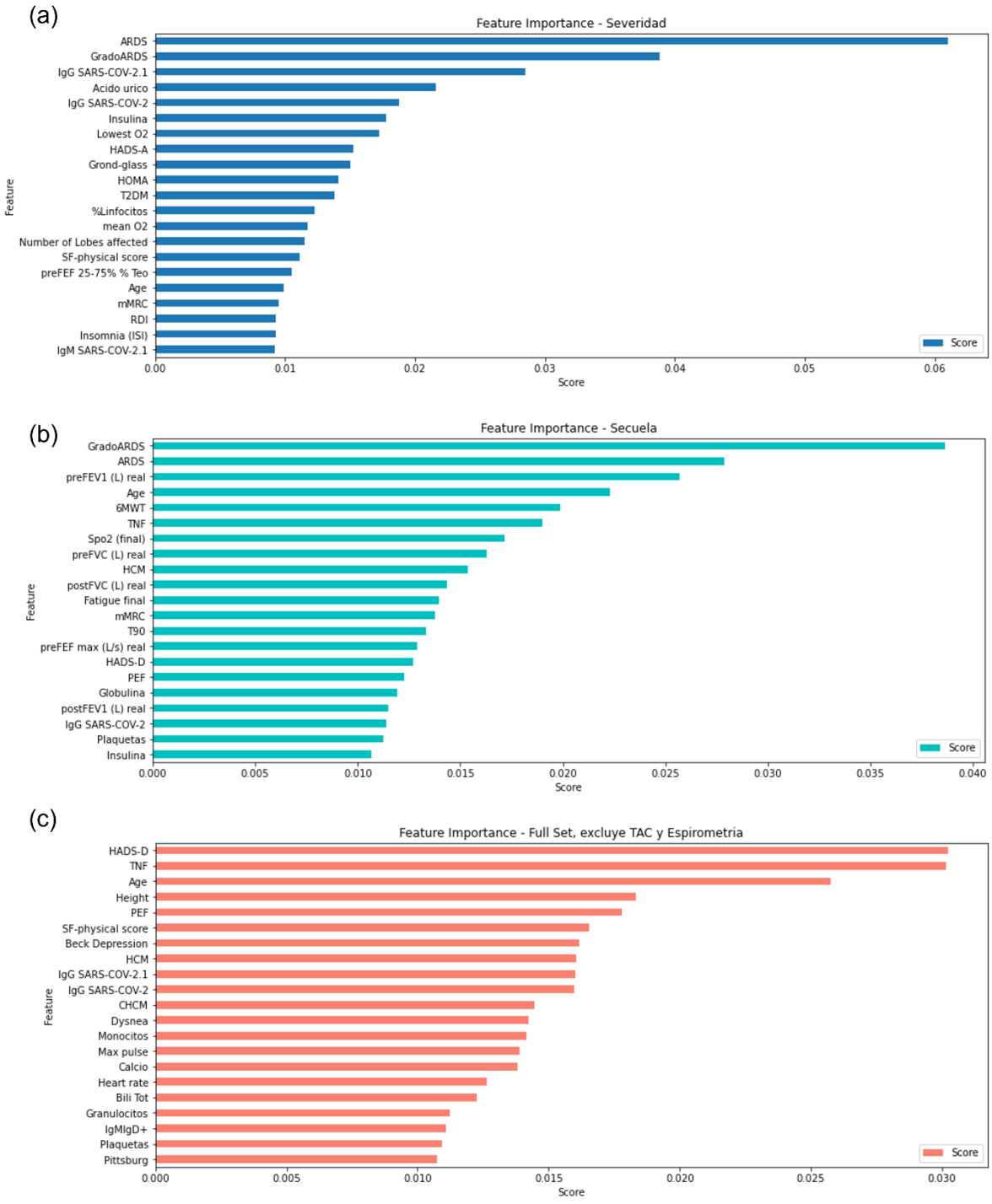

Ilustración 3. Gráfico Feature Importance - Set Completo. Se presentan en el eje X el valor de score asignado por el algoritmo a cada característica descrita en el eje Y. (a) Grado de severidad. (b) Nivel de secuela. (c) Nivel de secuela, excluye *TAC* y *Espirometría.*

11.2.2. Subsets

Actigrafía: a partir del set de datos que entregó el examen de actigrafía se obtuvo que las características más importantes para modelar *severidad* corresponden a mediciones asociadas a la estabilidad, ciclo y eficiencia del sueño. Estás son: *Intradaily Stability, Acrophase*, *SD Sleep Eff* y *CFI* (Ilustración 4a). En cuanto a las características más importantes para modelar secuela corresponden a: *Amplitude, Sleep Efficiency, Intradaily Stability,* y *SD Waso* (Ilustración 5a).

Clínica: las características más importantes obtenidas a partir del análisis de *feature importance* para grado de severidad en el subset de datos *Clínica* corresponden a parámetros metabólicos, destacando: *HOMA*, *Bilirrubina directa*, *Ácido Úrico* e *Insulina* (Ilustración 4b). Mientras que para nivel de secuela se obtuvo que Insulina, Globulina, Bilirrubina directa y Fosfatasa Alcalina corresponden a las características más importantes del modelo (Ilustración 5b)).

Cuestionario: de los datos obtenidos desde los cuestionarios se obtuvo que las características que más influyen en el modelo de grado de severidad corresponden a cuestionarios asociados al estrés físico y mental de los pacientes consultados. Las variables mejor puntuadas fueron *SF-physical score*, *Stop bang*, *HADS-A* y *SF-12 mental score* (Ilustración 4c). Por otra parte, para nivel de secuela, las variables mejor puntuadas corresponden a *HADS-A, SF-physical score, Pittsburg* y *SF-12 mental score (Ilustración 5c).*

Demografía: la evaluación de importancia de características en este *subset* entregó como resultado que las mediciones antropométricas son aquellas que aportan en mayor medida tanto al modelo de grado de severidad como al de nivel de secuela. Las características que destacan en ambos casos son: *índice de masa corporal*, *edad*, *peso* y *mediciones de perímetro corporal* (Ilustración 4d, Ilustración 5d).

Espirometría: de los datos obtenidos a partir de este examen se obtuvo que las características más importantes para el modelo corresponden a: *Spo2 (final y basal), postFEV(Volumen Espiratorio Forzado)*, *preFEF(Flujo espiratorio forzado)* y *postFVC(Capacidad Vital forzada)* (Ilustración 4e) para el caso de grado de

severidad, mientras que para evaluar el nivel de secuela se obtuvieron: *DLCO<80, Spo2 (final), postFVC (L) real* y *postFEV1 (L) real* como las variables más importantes (Ilustración 5e).

Hemograma: la evaluación de importancia de características para este análisis de sangre entregó como resultado que las características que entregan nivel y concentración de componentes sanguíneos serían las que más aportan al modelo de severidad. Estas características corresponden a: *Plaquetas*, *VCM (Volumen Corpuscular Medio)*, *%Linfocitos*, *HCM (Hemoglobina Corpuscular Media)* y *Granulocitos* (Ilustración 4f). Por otro lado, para el modelo de secuela se obtuvieron como las variables más importantes CHCM (Concentración Hemoglobina corpuscular media), HCM, Plaquetas y Conteo de Linfocitos (Ilustración 5f).

Poligrafía: En este subset las características que más aportan al modelo de severidad corresponden a mediciones de los valores de oxígeno, frecuencia respiratoria y pulso cardiaco durante el sueño. Las variables con más alto score fueron: *ODI (Desaturación de oxígeno)*, *RDI (Índice de trastorno respiratorio)* y *Pulso* (*mínimo*, *promedio*, *máximo*) (Ilustración 4g). Para el caso del modelo de secuela se obtuvieron que las variables mejor puntuadas por el algoritmo fueron: T90, Lowest O2, Mean O2, Max pulse y Min pulse (Ilustración 5g).

Proteína: en este set de datos la evaluación de características más importantes determinó que las tasas de *IL-8*, *B-cells*, *C4a* y *TNF* son las que más información entregan al modelo tanto para severidad como para secuela (Ilustración 4h, Ilustración 5h).

Sintomatología: las características más importantes, para modelar grado de severidad, obtenidas a partir de este subset corresponden a *Fiebre*, *Dolor de pecho*, *polipnea* y *preEuroQol (Ilustración 4i)*, mientras que para evaluar nivel de secuela se obtuvieron *preEuroQol, dolor de garganta, disnea* y *polipnea (Ilustración 5i)*.

Sintomatología 2: En este caso se obtuvo que la característica que más aportó al modelo severidad y secuela corresponde a *postEuroQol*, seguido de características

como *disnea*, *dolor de cabeza y tos para los modelos* de severidad y secuela (Ilustración 4j, Ilustración 5j).

TAC: la evaluación de importancia de características a partir del examen de tomografía axial computarizada entrega como resultado que variables como TSS total (Total Severity Score), Number of lobules affected, Ground-glass y Abnormal Chest-CT son aquellas que más información aportan a ambos modelos de clasificación (Ilustración 4k, Ilustración 5k).

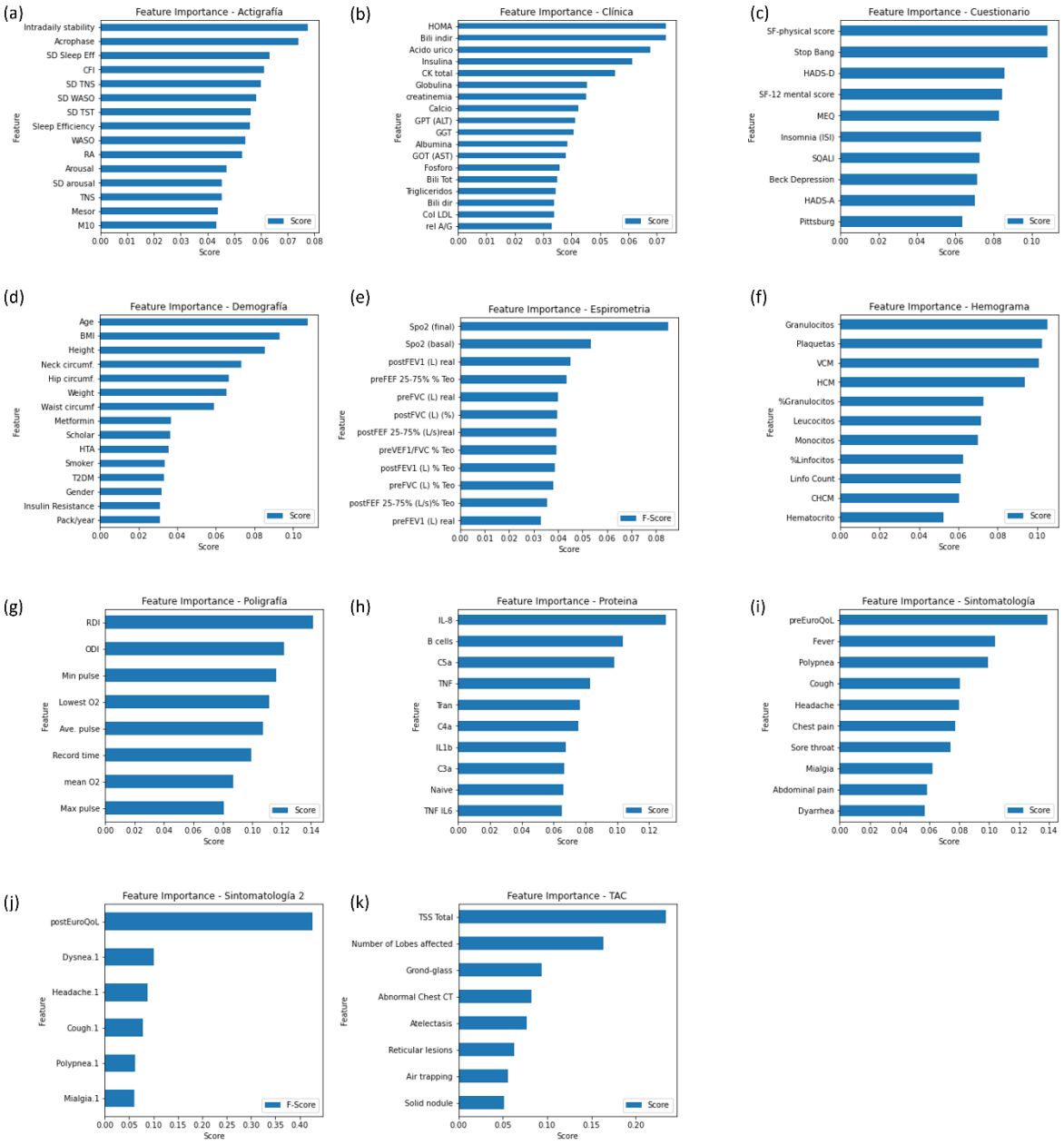

Ilustración 4. *Feature importance subsets de datos para grado de severidad**.** Se presentan en el eje X el valor de score asignado por el algoritmo a cada característica descrita en el eje Y. (a) Subset Actigrafía (b) Subset Clínica (c) Subset Cuestionario (d) Subset Demografía (e) Subset Espirometría (f) Subset Hemograma (g) Subset Poligrafía (h) Subset Proteína (i) Subset Sintomatología (j) Subset Sintomatologia2 (k) Subset TAC.*

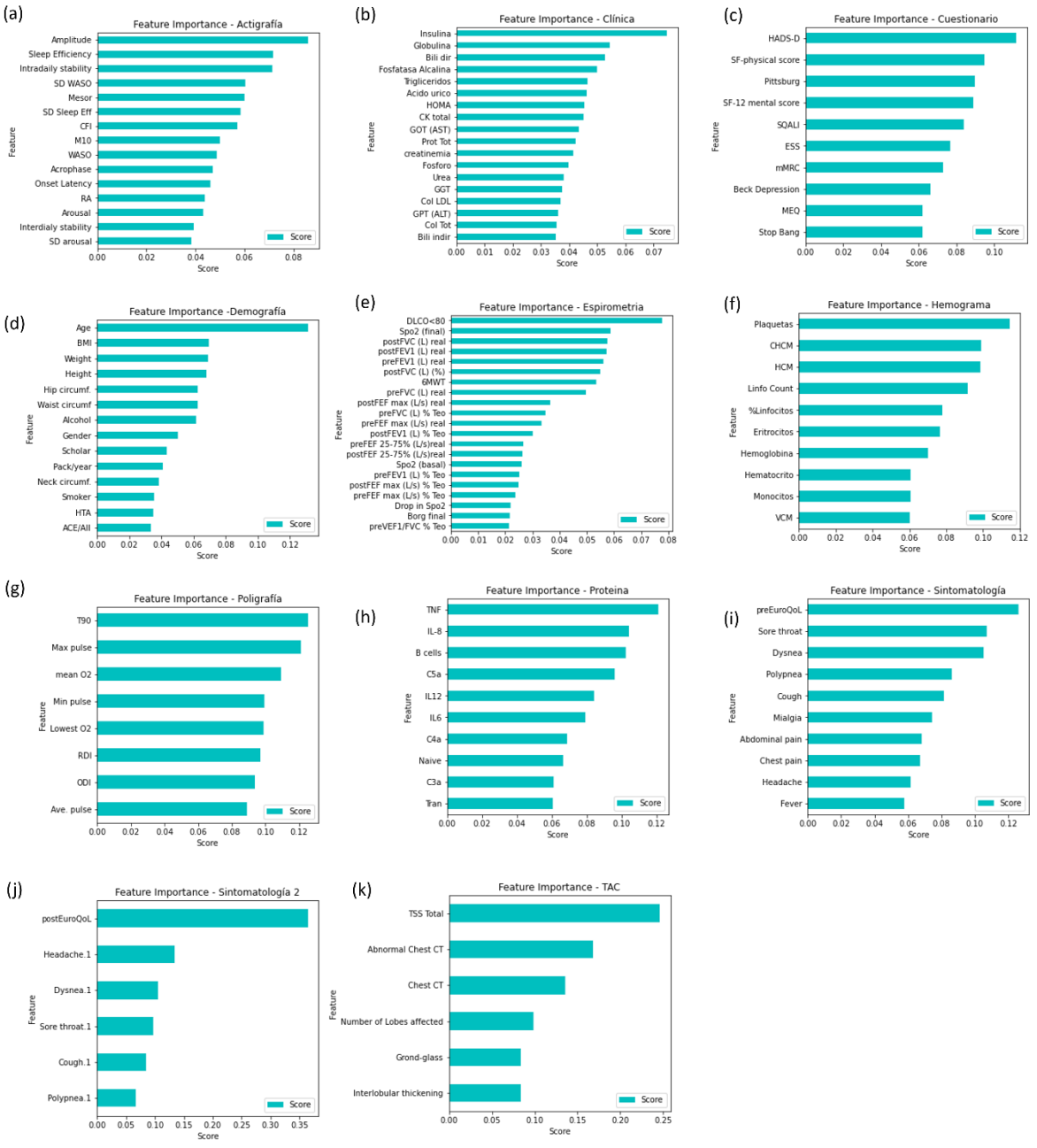

Ilustración 5. *Feature importance subsets de datos para nivel de secuela**.** Se presentan en el eje X el valor de score asignado por el algoritmo a cada característica descrita en el eje Y. (a) Subset Actigrafía (b) Subset Clínica (c) Subset Cuestionario (d) Subset Demografía (e) Subset Espirometría (f) Subset Hemograma (g) Subset Poligrafía (h) Subset Proteína (i) Subset Sintomatología (j) Subset Sintomatologia2 (k) Subset TAC.*

A partir de los resultados obtenidos por cada evaluación se generaron nuevos sets de datos incorporando solo aquellas características que aportaban en mayor medida a cada modelo propuesto. Estos sets de datos cuentan con una dimensionalidad reducida en comparación a cada subset original, de acuerdo con los resultados obtenidos de cada algoritmo de selección de características. La cantidad de características por subset de datos se muestran en la Tabla 7 para grado de severidad y nivel de secuela.

Tabla 7. *Número de características reducidas por cada algoritmo de selección de características*

Subset de datos	**Número de características**						
		Feature Importance		**Mutual Information**		**ANOVA**	
	Inicial	**Severidad**	**Secuela**	**Severidad**	**Secuela**	**Severidad**	**Secuela**
Set Completo	327	90	83	93	89	34	60
TAC	16	8	7	11	13	15	16
Clínica	25	17	18	9	7	8	13
Hemograma	14	10	10	10	10	7	13
Poligrafía	12	8	8	10	10	8	11
Cuestionario	14	10	10	11	11	12	12
Espirometría	33	21	20	13	10	26	31
Actigrafía	21	15	16	9	14	8	9
Demografía	35	13	15	13	9	25	10
Sintomatología	14	10	11	10	10	11	13
Sintomatología2	14	6	10	7	11	8	12

Adicionalmente se obtuvieron gráficos de SHAP con el fin de generar una representación visual de la forma en que el modelo se ajustaba a las variables que se le entregaron. Estos esquematizan el cómo cada característica aporta a la creación del modelo por cada clase objetivo, expuesto en la Ilustración 6.

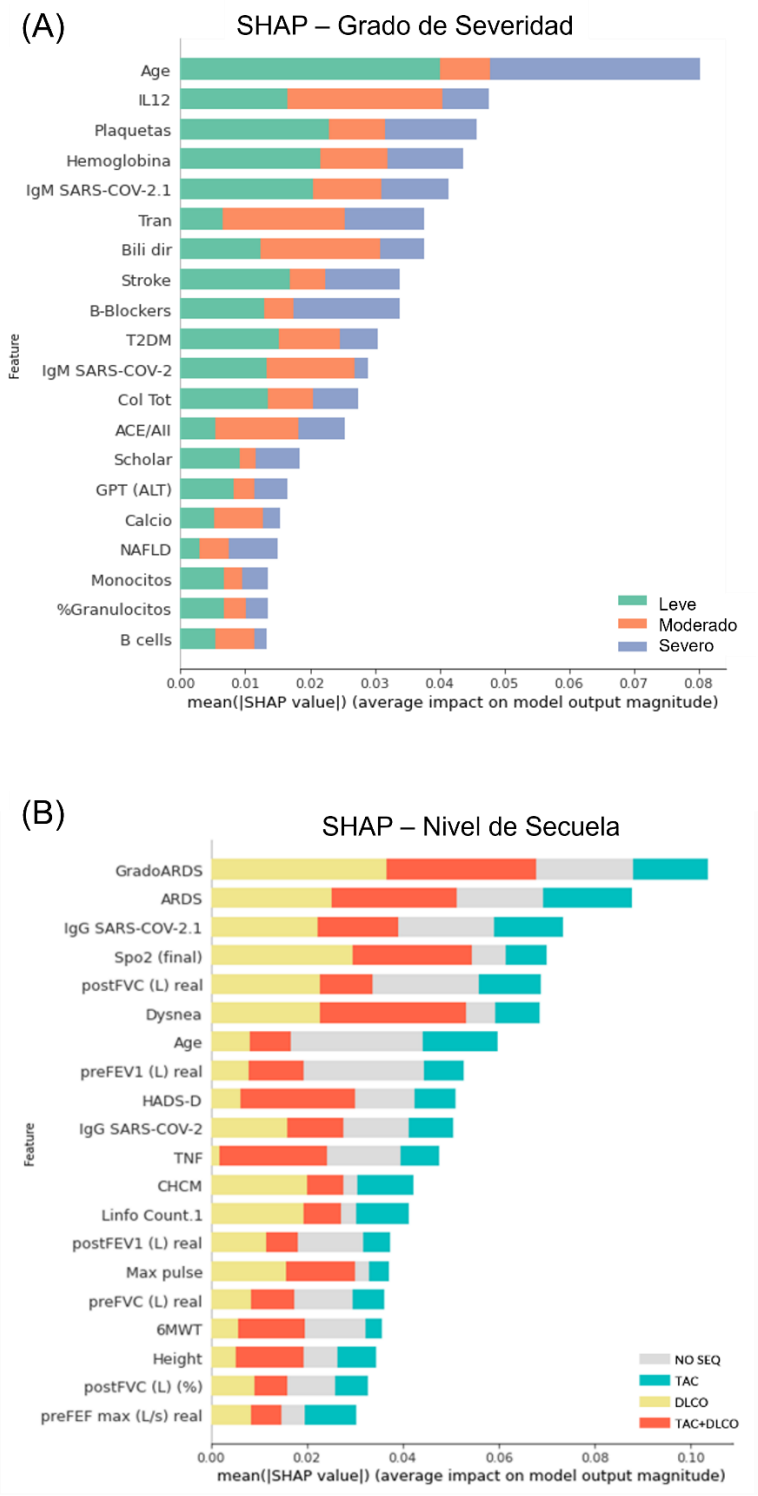

Ilustración 6. *Gráfico de SHAP – Set Completo. Se presenta en el eje X el SHAP-value asignado por el algoritmo a cada característica descrita en el eje Y. Los colores denotan el aporte de cada característica por clase en la generación del modelo. (A) SHAP – Grado de Severidad. Verde = Leve, Naranja = Moderado, Lila = Severo. (B) SHAP – Nivel de Secuela. Gris = No secuela, Turquesa = TAC alterado, Oro = DLCO alterado, Rojo = TAC y DLCO alterados.*

11.3. Resultados objetivo específico 3: Evaluar diferentes modelos de Inteligencia Artificial para clasificar grado de severidad y nivel de secuela en pacientes que cursaron COVID-19.

A partir de los modelos generados por cada método de clasificación se utilizó el algoritmo de validación cruzada (LOOCV) para evaluar el rendimiento de estos. Finalmente se implementó la obtención de métricas *accuracy*, *precision*, *recall* y *f1-score* para la validación de los modelos de clasificación generados.

11.3.1. Grado de severidad

En esta sección se exponen los dos sets de datos (*Set Completo*, *subset TAC*) en los que se obtuvo mejor rendimiento al momento de clasificar grado de severidad. Se describen los resultados para ambos casos a continuación:

Set Completo

Los resultados obtenidos al utilizar el método de *RandomForest,* en conjunto con el algoritmo de búsqueda de hiperparametros *RandomizedSearchCV*, lograron entregar un modelo de clasificación con una precisión del 89.3%, para el grado de severidad (Tabla 8, Ilustración 7), esto al evaluar el método con el set completo, con el subset generado a partir del algoritmo de selección de características *Feature Importance*. Adicionalmente se expone el gráfico de curvas ROC de los modelos expuestos en la tabla anteriormente mencionada y las métricas de *accuracy*, *precision*, *recall* y *f1-score* para cada clase en el modelo mejor evaluado por *LOOCV* (Tabla 9).

Tabla 8. *LOOCV accuracy – Set Completo, modelo de clasificación Random Forest para grado de severidad*

LOOCV – Random Forest – Severidad				
	FullSet	**FullSet_FI**	**FullSet_MI**	**FullSet_ANOVA**
Random Forest	0.787	0.773	0.840	0.747
GridSearchCV	0.787	0.827	0.867	0.827
RandomizedSearchCV	0.827	0.893	0.813	0.853

Tabla 9. *Métricas de evaluación por clase*

Random Forest – Set completo – Subset Feature Importance – RandomizedSearchCV				
Clase	***accuracy***	***precision***	***recall***	***f1-score***
Leve	1	1	1	1
Moderado	0.83	0.62	0.83	0.71
Severo	0.7	0.88	0.7	0.78

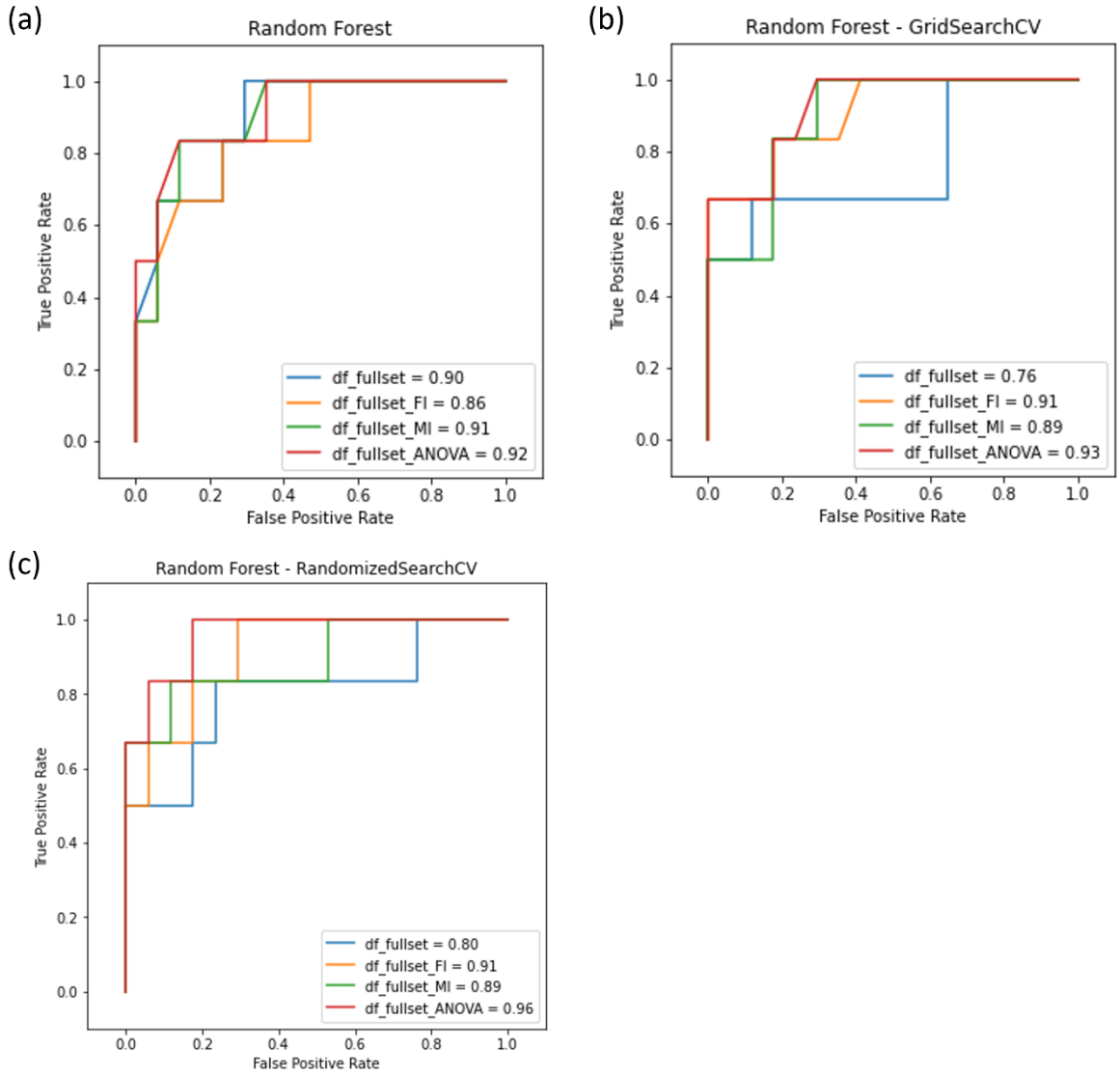

Ilustración 7. *Gráfico de curvas ROC, modelo de clasificación Random Forest para grado de severidad, Set Completo. En el eje X se presenta la tasa de falsos positivos y en el eje Y la tasa de Verdaderos Positivos. (a) Modelo Random Forest (b) Modelo Random Forest, GridSearchCV (c) Modelo Random Forest, RandomizedSearchCV.*

Por su parte, los modelos implementados con el algoritmo XGBoost obtuvieron puntuaciones superiores a 80% de precisión para todos los subsets generados a partir de los métodos de selección de características (Tabla 10, Ilustración 8)**,** se incluyen los valores de las métricas de evaluación por clase del modelo mejor puntuado en la Tabla 11.

Tabla 10. *LOOCV accuracy. Set completo, modelo de clasificación XGBoost, grado de severidad*

LOOCV – XGBoost – Severidad				
	FullSet	**FullSet_FI**	**FullSet_MI**	**FullSet_ANOVA**
XGBoost	0.827	0.853	0.813	0.813
GridSearchCV	0.787	0.840	0.840	0.827
RandomizedSearchCV	0.787	0.867	0.8	0.853

Tabla 11. *Métricas de evaluación por clase*

XGBoost – Set completo – Subset Feature Importance – RandomizedSearchCV				
Clase	***accuracy***	***precision***	***recall***	***f1-score***
Leve	0.857143	1	0.86	0.92
Moderado	0.833333	0.62	0.83	0.71
Severo	0.8	0.89	0.8	0.84

Subset TAC

A partir de la implementación de los modelos de clasificación para cada subset de datos se logró obtener un modelo de clasificación de grado de severidad a partir del set de datos *TAC*. Este logró un desempeño de un 80% de precisión al ser implementado con el método XGBoost (Tabla 12, Ilustración 9). También se incluyen las métricas de evaluación para este modelo en la Tabla 13.

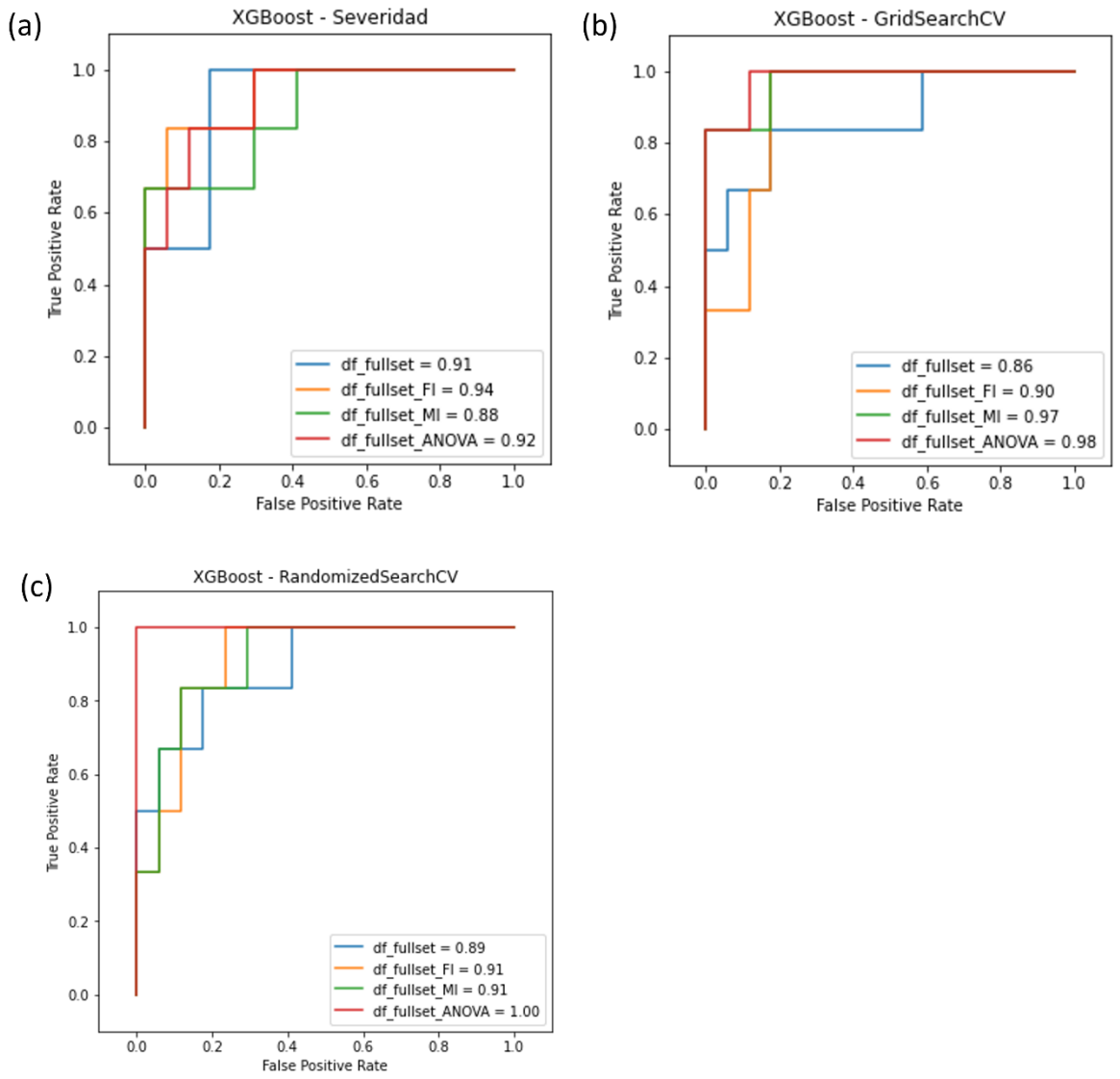

Ilustración 8. *Gráfico de curvas ROC modelo de clasificación XGBoost para grado de severidad, Set completo. En el eje X se presenta la tasa de falsos positivos y en el eje Y la tasa de Verdaderos Positivos. (a) Modelo XGBoost (b) Modelo XGBoost, GridSearchCV (c) Modelo XGBoost, RandomizedSearchCV.*

Tabla 12. *LOOCV accuracy – Set TAC, modelo de clasificación XGBoost para grado de severidad*

LOOCV – XGBoost – Severidad				
	TAC	**TAC_FI**	**TAC_MI**	**TAC_ANOVA**
XGBoost	0.8	0.56	0.587	0.547
GridSearchCV	0.8	0.613	0.6	0.52
RandomizedSearchCV	0.773	0.533	0.547	0.547

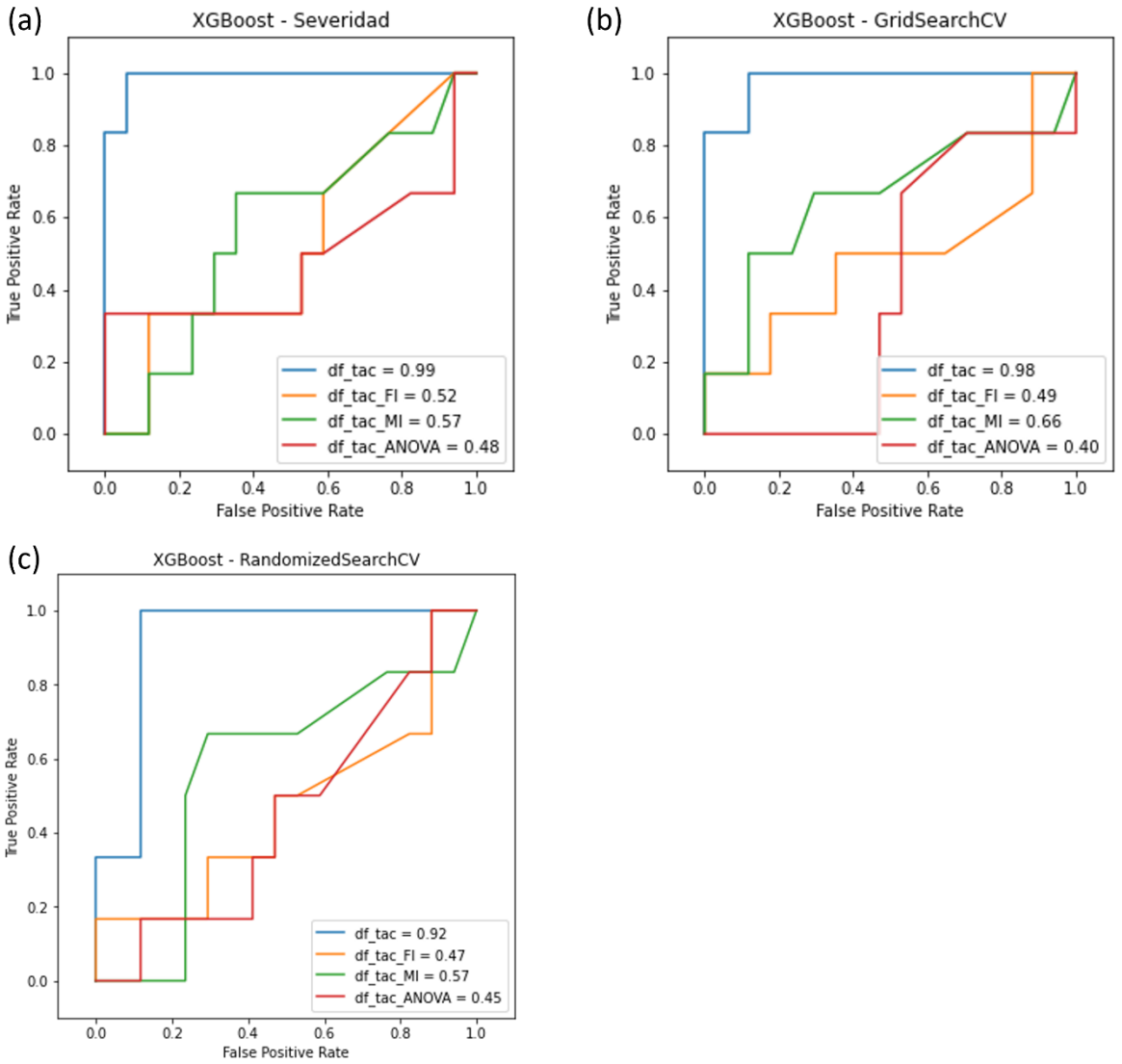

Ilustración 9. *Gráfico de curvas ROC modelo de clasificación XGBoost para grado de severidad, subset TAC. En el eje X se presenta la tasa de falsos positivos y en el eje Y la tasa de Verdaderos Positivos. (a) Modelo XGBoost (b) Modelo XGBoost, GridSearchCV (c) Modelo XGBoost, RandomizedSearchCV*

Tabla 13. *Métricas de evaluación modelo de clasificación de secuela*

XGBoost – Set TAC				
Clase	***accuracy***	***precision***	***recall***	***f1-score***
Leve	0.714286	1	0.71	0.83
Moderado	1	0.67	1	0.8
Severo	0.9	1	0.9	0.95

Nivel de secuela

En esta sección se exponen los resultados obtenidos de los dos sets de datos (*Set Completo*, *subset espirometría*) en los que se obtuvo mejor rendimiento al momento de clasificar nivel de secuela. Adicionalmente se exponen los resultados obtenidos al implementar el modelo de clasificación de secuela a partir del set de datos completo excluyendo las características de *TAC* y *Espirometría*. Se describen los resultados para los casos mencionados a continuación:

Set Completo

De los modelos de clasificación implementados para medir el nivel de secuela, el mejor resultado se obtuvo a través del método *Random Forest*, en conjunto con el algoritmo de búsqueda de hiperparámetros *GridSearchCV*. Su desempeño alcanzó una precisión del 100% para el set completo y el subset generado a partir del algoritmo de selección de características *Feature Importance* (Tabla 14, Ilustración 10). En este caso se observa que tanto el subset completo como los obtenidos a partir de los algoritmos de selección de características obtuvieron un gran desempeño a partir de los resultados expuestos por las métricas de evaluación (Tabla 15). Estos valores se repitieron para los tres modelos evaluados con precisión de 100%.

Tabla 14. *LOOCV accuracy – Set Completo, modelo de clasificación Random Forest para nivel de secuela*

LOOCV – Random Forest – Secuela				
	FullSet	**FullSet_FI**	**FullSet_MI**	**FullSet_ANOVA**
Random Forest	0.946	0.957	0.967	0.957
GridSearchCV	1	1	0.989	0.728
RandomizedSearchCV	1	0.989	0.989	0.750

Tabla 15. *Métricas de evaluación modelos de clasificación de secuela*

Random Forest – Set completo - Secuela				
Clase	***accuracy***	***precision***	***recall***	***f1-score***
NO SEQ	1	1	1	1
TAC ALTERADO	1	1	1	1
DLCO ALTERADO	1	1	1	1
TAC + DLCO ALTERADOS	1	1	1	1

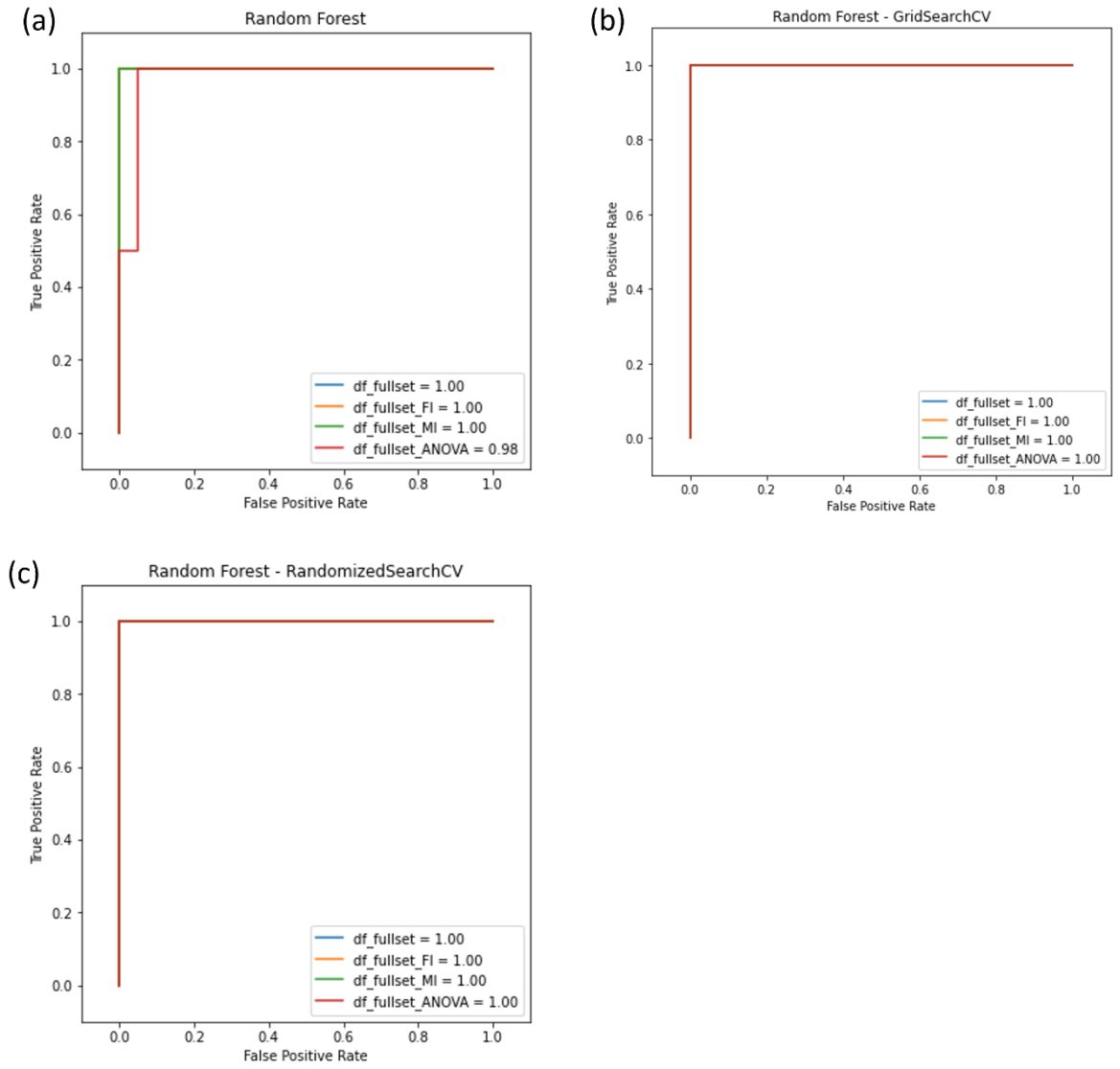

Ilustración 10. *Gráfico de curvas ROC modelo de clasificación Random Forest para nivel de secuela. En el eje X se presenta la tasa de falsos positivos y en el eje Y la tasa de Verdaderos Positivos. (a) Modelo Random Forest (b) Modelo Random Forest, GridSearchCV (c) Modelo Random Forest, RandomizedSearchCV.*

Set Espirometría

Otro set de datos que tuvo un buen rendimiento a partir de la evaluación de las métricas obtenidas fue el de espirometría. Este set de datos obtuvo una precisión superior al 80% al generar el modelo de clasificación de nivel de secuela. Este resultado se obtuvo con los modelos de *Random Forest* (Tabla 16, Ilustración 11) y *XGBoost* (Tabla 17, Ilustración 12). Adicionalmente se exponen en las métricas de evaluación por clase de los dos modelos mejor evaluados: *Random Forest – Mutual Information* (Tabla 18) y *XGBoost – Mutual Information* (Tabla 19).

Tabla 16. *LOOCV accuracy – SubSet Espirometria, modelo de clasificación Random Forest para nivel de secuela*

LOOCV – Random Forest – Secuela - Espirometria				
	espirometria	**espirometria_FI**	**espirometria_MI**	**espirometria_ANOVA**
Random Forest	0.793	0.826	0.870	0.793
GridSearchCV	0.750	0.772	0.826	0.728
RandomizedSearchCV	0.739	0.793	0.793	0.739

Tabla 17*. LOOCV accuracy – SubSet Espirometria, modelo de clasificación XGBoost para nivel de secuela*

LOOCV – XGBoost – Secuela - Espirometria				
	espirometria	**espirometria_FI**	**espirometria_MI**	**espirometria_ANOVA**
XGBoost	0.772	0.761	0.837	0.761
GridSearchCV	0.75	0.783	0.793	0.750
RandomizedSearchCV	0.826	0.804	0.826	0.783

Tabla 18. *Métricas de evaluación modelos de clasificación de secuela*

Random Forest – Set espirometría – Subset Mutual Information				
Clase	***accuracy***	***precision***	***recall***	***f1-score***
NO SEQ	0.71	0.50	0.71	0.59
TAC ALTERADO	1.00	0.89	1.00	0.94
DLCO ALTERADO	0.50	0.71	0.50	0.59
TAC + DLCO ALTERADOS	0.67	1.00	0.67	0.80

Tabla 19. *Métricas de evaluación modelos de clasificación de secuela*

XGBoost – Set espirometría – Subset Mutual Information				
Clase	***accuracy***	***precision***	***recall***	***f1-score***
NO SEQ	0.86	0.50	0.86	0.63
TAC ALTERADO	1.00	1.00	1.00	1.00
DLCO ALTERADO	5.00	0.83	0.5	0.62
TAC + DLCO ALTERADOS	0.67	1.00	0.67	0.80

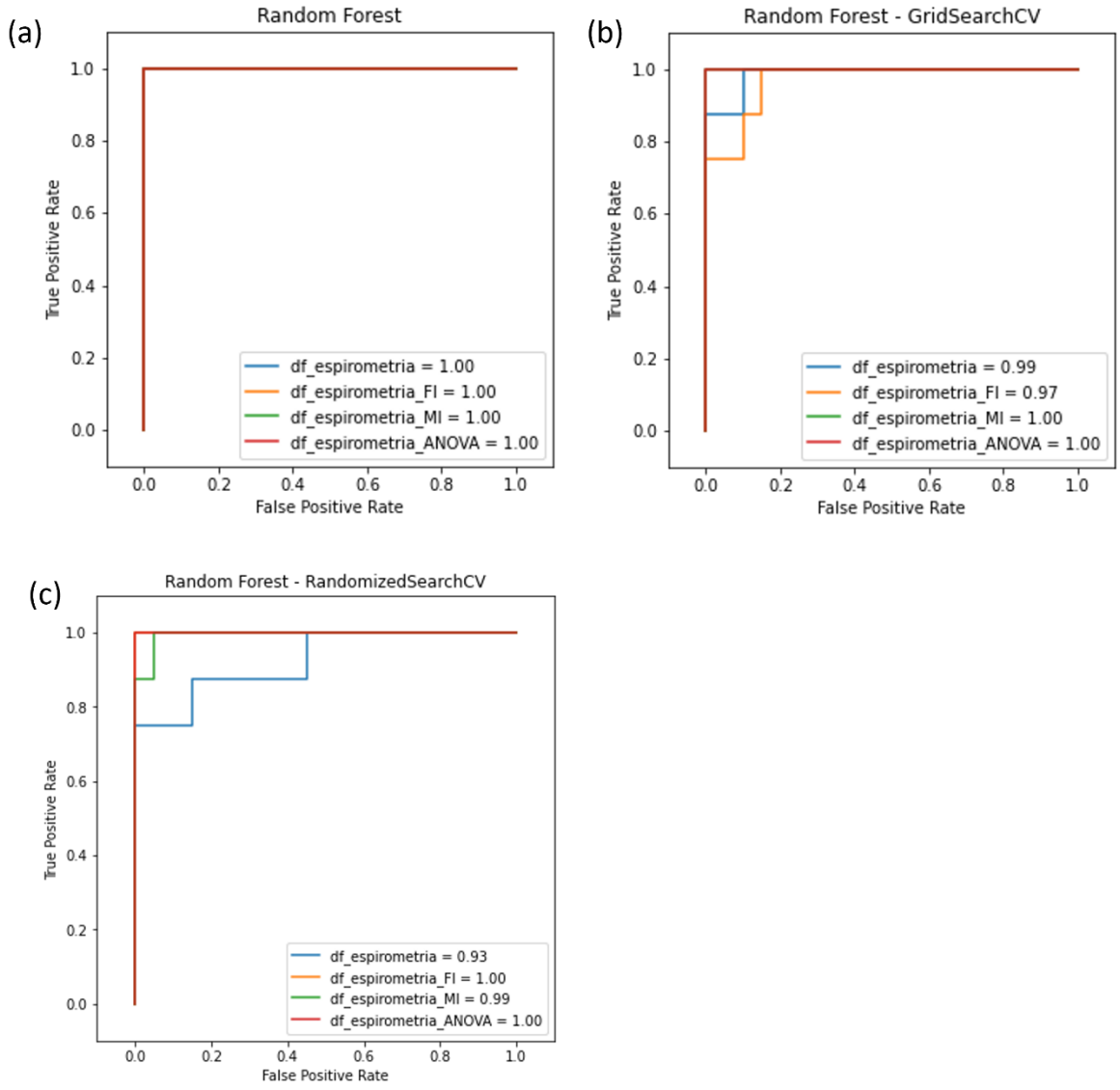

Ilustración 11. *Gráfico de curvas ROC modelo de clasificación Random Forest para nivel de secuela, Subset espirometria. En el eje X se presenta la tasa de falsos positivos y en el eje Y la tasa de Verdaderos Positivos. (a) Modelo Random Forest (b) Modelo Random Forest, GridSearchCV (c) Modelo Random Forest, RandomizedSearchCV.*

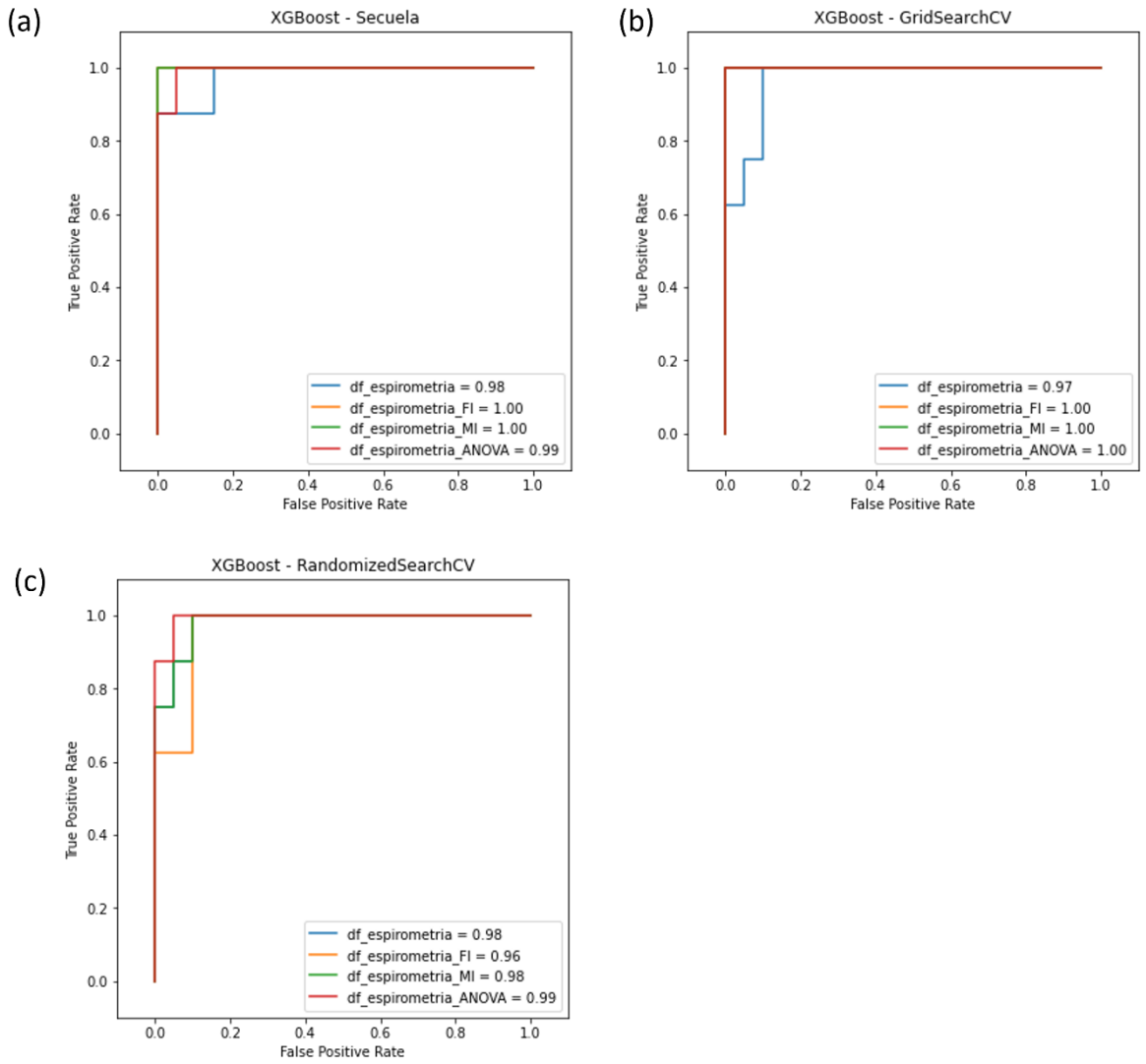

Ilustración 12. *Gráfico de curvas ROC modelo de clasificación XGBoost para nivel de secuela, Subset espirometria. En el eje X se presenta la tasa de falsos positivos y en el eje Y la tasa de Verdaderos Positivos. (a) Modelo XGBoost (b) Modelo XGBoost, GridSearchCV (c) Modelo XGBoost, RandomizedSearchCV.*

Set Completo – Excluido TAC y espirometría

Finalmente, se evaluaron los modelos de clasificación a partir del set de datos generado mediante la exclusión de las características de *TAC* y *espirometría* desde el *Set Completo*. Este subset de datos entregó como mejor resultado al algoritmo *RandomForest* en conjunto con el algoritmo de hiperparámetros *GridSearchCV* a partir del subset *ANOVA*, obteniendo un valor de precisión de *LOOCV–accuracy* de 75% (Tabla 20).

Tabla 20. *LOOCV accuracy – SubSet Completo, excluye TAC y Espirometria, modelo de clasificación Random Forest para nivel de secuela*

LOOCV – Random Forest– Secuela – Set Completo Excluye TAC y Espirometria - GridSearchCV				
	Subset_Completo	**Subset_FI**	**Subset_MI**	**Subset_ANOVA**
RandomForest	0.707	0.739	0.728	0.728
GridSearchCV	0.707	0.707	0.717	0.75
RandomizedSearchCV	0.696	0.652	0.652	0.674

También se exponen en la Tabla 21 las métricas de evaluación por clase del modelo mencionado anteriormente.

Tabla 21. *Métricas de evaluación modelos de clasificación de secuela*

Random Forest – Set Completo, Excluye TAC y Espirometria – Subset ANOVA - GridSearchCV				
Clase	***accuracy***	***precision***	***recall***	***f1-score***
NO SEQ	0.86	0.50	0.86	0.63
TAC ALTERADO	1.00	1.00	1.00	1.00
DLCO ALTERADO	5.00	0.83	0.5	0.62
TAC + DLCO ALTERADOS	0.67	1.00	0.67	0.80

La información entregada por las métricas de evaluación de los modelos de clasificación implementados permitió analizar de manera objetiva la precisión con la que los diferentes métodos generaban el proceso de clasificación. Se observó además que los valores de estas métricas alcanzaron valores más altos de *LOOCV-accuracy, accuracy, precision, recall* y *f1-score* en el caso de modelar el nivel de secuela, comparados con el grado de severidad, en donde se obtuvieron modelos con rendimientos inferiores.

12. DISCUSIÓN

12.1. Discusión resultados objetivo específico 1: Configurar un set de datos a partir de información clínica, experimental y demográfica obtenida desde 60 pacientes que cursaron COVID-19.

La información obtenida por el estudio observacional y cohorte prospectivo COVID-1005 significó el punto inicial para este estudio, siendo la base de la cual se construyeron, evaluaron y validaron distintos procesos de inteligencia artificial en busca de información que no estaba implícita en los registros.

Los datos obtenidos desde los 73 pacientes participantes de este estudio contaban con una alta variedad de características obtenidas desde distintos exámenes, ensayos clínicos, mediciones y cuestionarios. Dado esto, se realizó un *parseo* de la información a modo exploratorio, intentando indagar en la posible construcción de un modelo de clasificación a partir de un set más reducido, con características de un estudio o serie de estudios.

Cuando se trabajó inicialmente con los datos se encontraron limitaciones de acuerdo con la cantidad de información recolectada desde los 13 controles sanos, además de una serie de características sin información en algunos pacientes que, por distintas razones, dieron término a su seguimiento. Debido a estas razones se decidió descartar la información obtenida desde los pacientes control y de variables cuya información no estuviese completa. De esta forma se obtuvo el set de datos completo y los *subsets* definidos en la **Tabla 5.**

Las tareas realizadas en este objetivo permitieron establecer una base sólida para continuar con el trabajo planteado en la realización este estudio, logrando asegurar la correcta ejecución tanto de los algoritmos de selección de características como de la implementación y evaluación de los modelos de clasificación.

12.2. Discusión resultados objetivo específico 2: Determinar las características más relevantes asociadas a los diferentes grados de severidad y nivel de secuela en pacientes que cursaron COVID-19.

Al evaluar la información obtenida a partir de los algoritmos de importancia de características (Ilustración 3) se obtuvo que las variables *ARDS* (Síndrome de distrés respiratorio agudo) y *GradoARDS* son aquellas que aportan en mayor medida al modelo. Este resultado hace sentido debido a la importancia que tiene este parámetro al momento de determinar la severidad de un paciente. De la misma forma características como conteos de anticuerpos SARS-CoV-2 (*IgG SARS-CoV-2.1*, *IgG SARS-CoV-2*) forman parte importante para determinar nivel de severidad debido a su relación a una respuesta sistémica ante la infección con el patógeno.[76]

Por otra parte, se observó que las características asociadas a estudios cardiacos y pulmonares son las variables de mayor relevancia para los modelos de clasificación de severidad y secuela. El equipo de investigación esperaba que resultaran probables de obtener estas variables, debido a la relación existente entre las enfermedades cardiacas y pulmonares asociadas como factores de riesgo al cursar COVID-19 [25–27].

Otra característica asociada a una enfermedad que contribuye a la severidad y secuela es la *insulina*. Esta medición está presente como característica importante tanto en los modelos de severidad y secuela obtenidos a partir del set completo, como en los *subsets clínica* para ambos casos, demostrando así su relación con las dos clases objetivo. Dentro de la literatura se observa que existen mecanismos que relacionan la resistencia a la insulina con la severidad del COVID-19 [77] y la cohorte en estudio en esta investigación lo confirma. Dentro de este mismo set de características es que destacan otras tres variables: *Bilirrubina directa*, *ácido úrico* y *Globulina*, siendo las dos primeras objetos de estudio en cuanto a los niveles que presenta un paciente contagiado con el virus, logrando determinar su asociación directa con el grado de severidad y secuela [78–81], y la tercera siendo asociada como un posible tratamiento para concentrar la actividad neutralizadora de los anticuerpos anti SARS-CoV-2 a partir de plasma de individuos convalecientes [82], respaldando

los resultados obtenidos por el análisis de importancia de características realizado para este subset de datos.

Adicionalmente, se observa que otras características relevantes para determinar el grado de severidad y nivel de secuela son aquellas obtenidas a partir de exámenes de sangre y mediciones de citocinas plasmáticas. Estos resultados se relacionan con lo expuesto en diversos estudios en la literatura, que han logrado asociar grado de severidad y nivel de secuela a estas variables sistémicas [83–85].

En el análisis de importancia de características para el modelo de clasificación de severidad utilizando el *Set Completo, excluyendo las variables de TAC y Espirometría* (Ilustración 3c), se observaron características asociadas a las ya mencionadas antes, como lo son variables de cuestionario, citoquinas plasmáticas, demográficas, en donde destaca la ausencia de las variables que corresponden a los dos exámenes excluidos de este set de datos.

Por su parte, para el análisis por *subset* de datos, destacan variables como *postEuroQol* como una de las mejor puntuadas para el *subset Sintomatología 2* (Ilustración 4j, Ilustración 5j). Esta característica corresponde a un instrumento que evalúa la calidad de vida del paciente, significando un valor importante para determinar tanto severidad como secuela dentro de su *subset* de datos. En este sentido se tienen también las variables asociadas a la respuesta anímica y la percepción de la enfermedad en el organismo. Características como *HADS-A* figuran como una de las mejor puntuadas tanto en el *subset cuestionario* (Ilustración 4c, ilustración 5c), como en el *set completo* (Ilustración 3) para modelar severidad y secuela. Esta asociación fue validada y coincide con un estudio que evaluó nivel de hormonas y ansiedad en pacientes infectados de SARS-CoV-2 [86], logrando establecer esta relación entre la salud mental y la vulnerabilidad del paciente.

Al analizar el *subset demografía* (Ilustración 4d, Ilustración 5d) se observó que las mediciones antropométricas son las características más importantes. Estas características de composición corporal ya fueron asociadas con un desarrollo en la

severidad de pacientes con COVID-19 [87], logrando validar la información obtenida a partir de los resultados de este análisis para este *subset*.

Una de las características también asociadas al grado de severidad y nivel de secuela es la *edad*. Esta variable relacionada al desarrollo de diversas otras complicaciones sistémicas es una característica presente en la determinación de la respuesta de individuo al curso de la enfermedad en su cuerpo.

Otro análisis realizado fue el de SHAP-Value. A partir de este estudio se observa cómo las variables, como la recién mencionada *edad,* son importantes para definir en grado de severidad de un paciente que cursa COVID-19, así como la posibilidad de desarrollar secuela (Ilustración 6). Otras variables relevantes en este estudio corresponden a *IL12*, *Tran*, *Plaquetas* y *Hemoglobina*, que entregan información entre cada clase para determinar el grado de severidad, y obtener mejores resultados en el modelo de clasificación. Por otra parte, características como *Dysnea*, *TNF (tumor necrosis factor)*, *GradoARDS*, *HADS-D* y *Spo2 (final)* corresponden a las variables más relevantes para el modelo de clasificación en cada clase de nivel de secuela.

12.3. Discusión resultados objetivo específico 3: Evaluar diferentes modelos de Inteligencia Artificial para clasificar grado de severidad y nivel de secuela en pacientes que cursaron COVID-19.

Los modelos de clasificación implementados con los métodos *Random Forest*, *XGBoost*, *Decision Tree*, *Naive Bayes* y *Support Vector Machine* fueron entrenados para cada set y subset de datos para las dos clases objetivo de severidad y secuela. Los resultados se evaluaron y validaron mediante el método LOOCV de validación cruzada para cada modelo de clasificación. El análisis de estos resultados se describe a continuación:

Al implementar los modelos de clasificación con el set de datos completo se observa que el algoritmo con mejor desempeño es el de *RandomForest*, seguido por *XGBoost*. Ambos algoritmos lograron generar exitosamente modelos de clasificación a partir de los datos suministrados, entregando un valor de *LOOCV-*

accuracy de hasta 100%. Se observa en general que los algoritmos basados en árboles de decisión, obtienen un mejor desempeño, disminuyendo el error y evaluando las mejores combinaciones de variables al momento de generar los nodos que relacionan las características de entrada con la clase objetivo [69,70].

Por su parte, el modelo de clasificación de nivel de secuela a partir del *Set Completo, excluyendo las variables de TAC y Espirometría* demostró un rendimiento inferior a 80% para validar la implementación del modelo. Este procedimiento se realizó a modo exploratorio, intentando excluir las variables del set de datos más representativos al momento de determinar el nivel de severidad de un paciente con COVID-19. De esta forma, y a partir de los resultados expuestos en la Tabla 20 se puede determinar tanto las variables de *TAC* como de *espirometría* son esenciales para lograr clasificar de manera precisa a los pacientes de acuerdo a su posibilidad de desarrollar secuela.

Por su parte, la implementación de modelos de clasificación en los *subset* de datos determinó que los *subset TAC* y *subset espirometría* son los subsets más relevantes para determinar el grado de severidad y nivel de secuela en los pacientes. Este resultado abre a la posibilidad de obtener una clasificación a partir de un número más reducido de características asociadas a un solo examen.

13. CONCLUSIÓN

En esta memoria de título se utilizaron datos clínicos, parámetros sistémicos y demográficos de 73 individuos pertenecientes a la región del Biobío. Se implementaron y evaluaron modelos de inteligencia artificial para abordar los objetivos planteados de obtener las características más importantes y clasificar los distintos grados de severidad y el nivel de secuela en pacientes que cursaron COVID-19.

La hipótesis planteada se confirma de acuerdo con los resultados obtenidos, logrando establecer una relación entre la información de los individuos y la manifestación sintomática del paciente durante y posterior al cuadro infeccioso de COVID-19. A partir de los modelos de clasificación generados por los métodos de *Random Forest* y *XGBoost,* validados al evaluar las métricas de *LOOCV-accuracy*, *accuracy*, *precisión*, *recall* y *f1-score*, se obtuvieron modelos con rendimientos superiores a un 80% de precisión para grado de severidad y nivel de secuela. Además, mediante el análisis de importancia de características se logró determinar que mediciones clínicas, de tomografía axial computarizada y espirometría son aquellas con los atributos más influyentes para determinar grado de severidad y nivel de secuela en pacientes que cursaron COVID-19.

Por lo tanto, a partir de este estudio fue posible aportar al análisis de información generado por la pandemia de COVID-19, logrando integrar la información obtenida desde pacientes que cursaron esta enfermedad, a modo de apoyo a las diferentes técnicas actualmente utilizadas, en este caso desde el enfoque de la inteligencia artificial y el aprendizaje automático.

14. BIBLIOGRAFÍA

1. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol* **17**, 181–192 (2019).
2. Luk, H. K. H., Li, X., Fung, J., Lau, S. K. P. & Woo, P. C. Y. Molecular epidemiology, evolution and phylogeny of SARS coronavirus. *Infect Genet Evol* **71**, 21–30 (2019).
3. Zhong, N. *et al.* Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *The Lancet* **362**, 1353–1358 (2003).
4. Velavan, T. P. & Meyer, C. G. The COVID-19 epidemic. *Trop Med Int Health* **25**, 278–280 (2020).
5. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
6. Sun, P., Lu, X., Xu, C., Sun, W. & Pan, B. Understanding of COVID-19 based on current evidence. *Journal of Medical Virology* **92**, 548–551 (2020).
7. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* **20**, 533–534 (2020).
8. Wang, M.-Y. *et al.* SARS-CoV-2: Structure, Biology, and Structure-Based Therapeutics Development. *Front. Cell. Infect. Microbiol.* **10**, 587269 (2020).
9. Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N Engl J Med* **382**, 1199–1207 (2020).
10. Hu, B., Guo, H., Zhou, P. & Shi, Z.-L. Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology* **19**, 141–154 (2021).

11. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China - The Lancet. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30183-5/fulltext.
12. Chen, T. *et al.* Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study. *BMJ* **368**, m1091 (2020).
13. Bhadra, S. *et al.* Real-Time Sequence-Validated Loop-Mediated Isothermal Amplification Assays for Detection of Middle East Respiratory Syndrome Coronavirus (MERS-CoV). *PLOS ONE* **10**, e0123126 (2015).
14. Chu, D. K. W. *et al.* Molecular Diagnosis of a Novel Coronavirus (2019-nCoV) Causing an Outbreak of Pneumonia. *Clinical Chemistry* **66**, 549–555 (2020).
15. Tang, Y.-W., Schmitz, J. E., Persing, D. H. & Stratton, C. W. Laboratory Diagnosis of COVID-19: Current Issues and Challenges. *Journal of Clinical Microbiology* **58**, (2020).
16. Russo, A. G., Decarli, A. & Valsecchi, M. G. Strategy to identify priority groups for COVID-19 vaccination: A population based cohort study. *Vaccine* (2021) doi:10.1016/j.vaccine.2021.03.076.
17. Tadic, M., Cuspidi, C., Grassi, G. & Mancia, G. COVID-19 and arterial hypertension: Hypothesis or evidence? *J Clin Hypertens (Greenwich)* (2020) doi:10.1111/jch.13925.
18. de Leeuw, A. J. M., Oude Luttikhuis, M. A. M., Wellen, A. C., Müller, C. & Calkhoven, C. F. Obesity and its impact on COVID-19. *J Mol Med* (2021) doi:10.1007/s00109-021-02072-4.
19. Yang, J., Hu, J. & Zhu, C. Obesity aggravates COVID-19: A systematic review and meta-analysis. *J Med Virol* **93**, 257–261 (2021).

20. Vecchié, A. *et al.* Obesity phenotypes and their paradoxical association with cardiovascular diseases. *European Journal of Internal Medicine* **48**, 6–17 (2018).
21. Albashir, A. A. D. The potential impacts of obesity on COVID-19. *Clin Med (Lond)* **20**, e109–e113 (2020).
22. Peric, S. & Stulnig, T. M. Diabetes and COVID-19. *Wien Klin Wochenschr* 1–6 (2020) doi:10.1007/s00508-020-01672-3.
23. Guo, W. *et al.* Diabetes is a risk factor for the progression and prognosis of COVID-19. *Diabetes/Metabolism Research and Reviews* **36**, e3319 (2020).
24. Chang, W.-T., Toh, H. S., Liao, C.-T. & Yu, W.-L. Cardiac Involvement of COVID-19: A Comprehensive Review. *Am J Med Sci* **361**, 14–22 (2021).
25. Manolis, A. S. *et al.* COVID-19 infection and cardiac arrhythmias. *Trends Cardiovasc Med* **30**, 451–460 (2020).
26. Babapoor-Farrokhran, S. *et al.* Myocardial injury and COVID-19: Possible mechanisms. *Life Sci* **253**, 117723 (2020).
27. Topol, E. J. COVID-19 can affect the heart. *Science* **370**, 408–409 (2020).
28. Alonso-Lana, S., Marquié, M., Ruiz, A. & Boada, M. Cognitive and Neuropsychiatric Manifestations of COVID-19 and Effects on Elderly Individuals With Dementia. *Front Aging Neurosci* **12**, (2020).
29. Cilia, R. *et al.* Effects of COVID -19 on Parkinson's Disease Clinical Features: A COMMUNITY-BASED CASE-CONTROL Study. *Mov Disord* **35**, 1287–1292 (2020).
30. Aluja-Banet, T. La minería de datos, entre la estadística y la inteligencia artificial. *Questiio: Quaderns d'Estadistica, Sistemes, Informatica i Investigació Operativa, ISSN 0210-8054, Vol. 25, Nº. 3, 2001, pags. 479-498* (2001).
31. Mena, J. *Data Mining Your Website*. (Butterworth-Heinemann, 1999).

32. Kaplan, A. & Haenlein, M. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons* **62**, 15–25 (2019).
33. Yu, K.-H., Beam, A. L. & Kohane, I. S. Artificial intelligence in healthcare. *Nat Biomed Eng* **2**, 719–731 (2018).
34. Parunak, H. V. D. *Applications of Distributed Artificial Intelligence in Industry*. (1994).
35. Jovic, A., Brkic, K. & Bogunovic, N. A review of feature selection methods with applications. in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* 1200–1205 (IEEE, 2015). doi:10.1109/MIPRO.2015.7160458.
36. Forman, G. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. 27.
37. Liu, T., Liu, S., Chen, Z. & Ma, W.-Y. An Evaluation on Feature Selection for Text Clustering. in 488–495 (2003).
38. Bins, J. & Draper, B. A. Feature selection from huge feature sets. in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001* vol. 2 159–165 (IEEE Comput. Soc, 2001).
39. Muštra, M., Grgić, P. M. & Delač, K. Breast Density Classification Using Multiple Feature Selection. *Automatika* **53**, 362–372 (2012).
40. Dessì, N., Pascariello, E. & Pes, B. A Comparative Analysis of Biomarker Selection Techniques. *BioMed Research International* **2013**, e387673 (2013).

41. Abusamra, H. A Comparative Study of Feature Selection and Classification Methods for Gene Expression Data of Glioma. *Procedia Computer Science* **23**, 5–14 (2013).
42. Liu, C., Jiang, D. & Yang, W. Global geometric similarity scheme for feature selection in fault diagnosis. *Expert Systems with Applications* **41**, 3585–3595 (2014).
43. Lip, G. Y. H., Nieuwlaat, R., Pisters, R., Lane, D. A. & Crijns, H. J. G. M. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest* **137**, 263–272 (2010).
44. Bishop, C. M. *Pattern recognition and machine learning*. (Springer, 2006).
45. Deo, R. C. Machine Learning in Medicine. *Circulation* **132**, 1920–1930 (2015).
46. Goudbeek, M., Swingley, D. & Smits, R. Supervised and unsupervised learning of multidimensional acoustic categories. *J Exp Psychol Hum Percept Perform* **35**, 1913–1933 (2009).
47. Ramesh, A. N., Kambhampati, C., Monson, J. R. T. & Drew, P. J. Artificial intelligence in medicine. *Ann R Coll Surg Engl* **86**, 334–338 (2004).
48. Johnson, K. W. *et al.* Artificial Intelligence in Cardiology. *J Am Coll Cardiol* **71**, 2668–2679 (2018).
49. Harrer, S., Shah, P., Antony, B. & Hu, J. Artificial Intelligence for Clinical Trial Design. *Trends Pharmacol Sci* **40**, 577–591 (2019).
50. Hessler, G. & Baringhaus, K.-H. Artificial Intelligence in Drug Design. *Molecules* **23**, (2018).

51. Yang, X., Wang, Y., Byrne, R., Schneider, G. & Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem Rev* **119**, 10520–10594 (2019).
52. Kira, K. & Rendell, L. A. A Practical Approach to Feature Selection. in *Machine Learning Proceedings 1992* 249–256 (Elsevier, 1992). doi:10.1016/B978-1-55860-247-2.50037-1.
53. Guyon, I. & Elisseeff, A. An Introduction to Variable and Feature Selection. 26.
54. Chandrashekar, G. & Sahin, F. A survey on feature selection methods. *Computers & Electrical Engineering* **40**, 16–28 (2014).
55. Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G. & Chatzisavvas, K. Ch. A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory* **55**, 1–9 (2015).
56. Datta, N. R. & Datta, S. Comprehensive analysis of the key epidemiological parameters to evaluate the impact of BCG vaccination on COVID-19 pandemic. *medRxiv* 2020.08.12.20173617 (2020) doi:10.1101/2020.08.12.20173617.
57. Russell, S. J. & Norvig, P. *Artificial intelligence: a modern approach*. (Prentice Hall, 1995).
58. *Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, No. 2). Cambridge: MIT press.*
59. Gangloff, C., Rafi, S., Bouzillé, G., Soulat, L. & Cuggia, M. Machine learning is the key to diagnose COVID-19: a proof-of-concept study. *Sci Rep* **11**, (2021).
60. Banerjee, A. *et al.* Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population. *International Immunopharmacology* **86**, 106705 (2020).

61. Hass, F. S. & Jokar Arsanjani, J. The Geography of the Covid-19 Pandemic: A Data-Driven Approach to Exploring Geographical Driving Forces. *International Journal of Environmental Research and Public Health* **18**, 2803 (2021).
62. Reback, J. *et al.* pandas-dev/pandas: Pandas 1.0.3. *Zenodo* (2020) doi:10.5281/zenodo.3509134.
63. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
64. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
65. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II*. vol. 5782 (Springer Berlin Heidelberg, 2009).
66. François, D., Wertz, V. & Verleysen, M. The permutation test for feature selection by mutual information. in 239–244 (2006).
67. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. 10.
68. Quinlan, J. R. Induction of decision trees. *Mach Learn* **1**, 81–106 (1986).
69. Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001).
70. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016). doi:10.1145/2939672.2939785.
71. Rish, I. An Empirical Study of the Naïve Bayes Classifier. *IJCAI 2001 Work Empir Methods Artif Intell* **3**, (2001).

72. Kotsiantis, S. B. Supervised Machine Learning: A Review of Classification Techniques. 20.
73. Fisher, R. A. Baye's theorem. *Eugen Rev* **18**, 32–33 (1926).
74. Xu, L. *et al.* Stochastic cross validation. *Chemometrics and Intelligent Laboratory Systems* **175**, 74–81 (2018).
75. Gunawardana, A. & Shani, G. A Survey of Accuracy Evaluation Metrics of Recommendation Tasks. 28.
76. Dehgani-Mobaraki, P., Zaidi, A. K., Porreca, A., Floridi, A. & Floridi, E. *Antibody persistency and trend post-SARS-CoV-2 infection at eight months*. 2020.11.21.20236117 https://www.medrxiv.org/content/10.1101/2020.11.21.20236117v2 (2020) doi:10.1101/2020.11.21.20236117.
77. Finucane, F. M. & Davenport, C. Coronavirus and Obesity: Could Insulin Resistance Mediate the Severity of Covid-19 Infection? *Frontiers in Public Health* **8**, 184 (2020).
78. Liu, Z. *et al.* Bilirubin Levels as Potential Indicators of Disease Severity in Coronavirus Disease Patients: A Retrospective Cohort Study. *Frontiers in Medicine* **7**, 799 (2020).
79. Paliogiannis, P. & Zinellu, A. Bilirubin levels in patients with mild and severe Covid-19: A pooled analysis. *Liver Int* 10.1111/liv.14477 (2020) doi:10.1111/liv.14477.
80. Li, G. *et al.* Uric acid as a prognostic factor and critical marker of COVID-19. *Sci Rep* **11**, 17791 (2021).

81. Hu, F. *et al.* Association of serum uric acid levels with COVID-19 severity. *BMC Endocrine Disorders* **21**, 97 (2021).
82. Vandeberg, P. *et al.* Production of anti-SARS-CoV-2 hyperimmune globulin from convalescent plasma. *Transfusion* **61**, 1705–1709 (2021).
83. Yao, H. *et al.* Severity Detection for the Coronavirus Disease 2019 (COVID-19) Patients Using a Machine Learning Model Based on the Blood and Urine Tests. *Frontiers in Cell and Developmental Biology* **8**, 683 (2020).
84. Alahmad, B., Al-Shammari, A. A., Bennakhi, A., Al-Mulla, F. & Ali, H. Fasting Blood Glucose and COVID-19 Severity: Nonlinearity Matters. *Diabetes Care* **43**, 3113–3116 (2020).
85. Rod, J. E., Oviedo-Trespalacios, O. & Cortes-Ramirez, J. A brief-review of the risk factors for covid-19 severity. *Rev. saúde pública* **54**, 60 (2020).
86. Ramezani, M. *et al.* The Role of Anxiety and Cortisol in Outcomes of Patients With Covid-19. *Basic Clin Neurosci* **11**, 179–184 (2020).
87. Poros, B. *et al.* Anthropometric analysis of body habitus and outcomes in critically ill COVID-19 patients. *Obesity Medicine* **25**, 100358 (2021).

15. ANEXO

1. Consentimiento informado

Comité Ético Científico
Servicio Salud Biobío

Nº 98

Ref.: Respuesta Solicitud de Aprobación Protocolo **"Neumonía Viral Secundaria a Corona Virus en el Complejo Asistencial Dr. Víctor Ríos Ruiz"**

Los Ángeles, 06 de Abril de 2020

A: Investigador Principal Sr. Gonzalo Labarca Trucios

DE: Comité Etico de Investigación

Estimado Investigador:

El Comité Ético Científico del Servicio de Salud Biobío, en su sesión ordinaria de fecha 26.03.2020 ha analizado los antecedentes remitidos por Ud. en relación a solicitud de aprobación protocolo en referencia, resolviendo lo siguiente:

1. **Valor del estudio:** La investigación aportara nuevos antecedentes sobre los atributos clínicos del COVID-19, perfiles demográficos de los pacientes con diagnóstico de neumonía, desenlaces a corto y largo plazo según nivel de severidad del cuadro. Acumulando evidencia objetiva, que a futuro contribuirá a reforzar las medidas de prevención primaria y de manejo posterior de pacientes hospitalizados con diagnóstico de neumonía secundario a Covid-19 en un establecimiento de salud pública a nivel local.

1. **Validez científica:** Estudio observacional, prospectivo de corte transversal.

2. **Discriminación arbitraria:** Se ajusta a criterios de inclusión y exclusión delimitados en el marco de la investigación.

3. **Relación riesgo-beneficio:** La probabilidad de que ocurra un efecto nocivo a consecuencia de la implementación de la investigación es baja.

4. **Consentimiento informado:** Se debe ajustar a lo indicado los artículos 8°, 10°, 14° y 15° de la Ley Nº 20.584 sobre consentimiento informado (CI) con sus exigencias y excepciones y el Derecho de las personas a decidir informadamente.

EN CONSECUENCIA,

En virtud de los antecedentes tenidos a la vista, y toda vez que el procedimiento puesto en marcha, se ajuste a los artículos mencionados en el punto 4, este Comité ha resuelto "aprobar" la realización del protocolo de investigación en referencia.

Tomó conocimiento de:

- ✓ Formulario Acceso CEC.
- ✓ Protocolo de Investigación.
- ✓ Consentimiento Informado.
- ✓ Carta de autorización de: Director Complejo Asistencial Dr. Víctor Ríos Ruiz.
- ✓ Currículo Vitae Investigador Principal.

Sin otro particular, les saluda muy atentamente,

CARLOS VILLARROEL INOSTROZA
PRESIDENTE
COMITÉ ÉTICO CIENTÍFICO

En respuesta a su solicitud el CEC (Comité Ético Científico) se reunió en sesión ordinaria con fecha 26.03.2020 estando presente Dr. Carlos Villarroel, A.S (MC) Patricia Messenger, E.U. Cecilia Alvarado y B.Q. Felipe Riquelme.

DISTRIBUCION:

-Interesado.
-Archivo

2. Identificadores de pacientes

Tabla Anexo I. *Identificadores correspondientes a los 73 individuos de estudio.*

ID			
COVID1005-1	COVID1005-21	COVID1005-41	Control_1
COVID1005-2	COVID1005-22	COVID1005-42	Control_2
COVID1005-3	COVID1005-23	COVID1005-43	Control_3
COVID1005-4	COVID1005-24	COVID1005-44	Control_4
COVID1005-5	COVID1005-25	COVID1005-45	Control_5
COVID1005-6	COVID1005-26	COVID1005-46	Control_6
COVID1005-7	COVID1005-27	COVID1005-47	Control_7
COVID1005-8	COVID1005-28	COVID1005-48	Control_8
COVID1005-9	COVID1005-29	COVID1005-49	Control_9
COVID1005-10	COVID1005-30	COVID1005-50	Control_10
COVID1005-11	COVID1005-31	COVID1005-51	Control_11
COVID1005-12	COVID1005-32	COVID1005-52	Control_12
COVID1005-13	COVID1005-33	COVID1005-53	Control_13
COVID1005-14	COVID1005-34	COVID1005-54	
COVID1005-15	COVID1005-35	COVID1005-55	
COVID1005-16	COVID1005-36	COVID1005-56	
COVID1005-17	COVID1005-37	COVID1005-57	
COVID1005-18	COVID1005-38	COVID1005-58	
COVID1005-19	COVID1005-39	COVID1005-59	
COVID1005-20	COVID1005-40	COVID1005-60	

3. Identificadores de características

Tabla Anexo II. *características contenidas en el set de datos*

ID				
Código muestra	GradoARDS	NAFLD	Dyarrhea	HADS-D
Grado	B cells	Hypothyroidism	Change smell	Beck Depression
Glicemia	CD27+	ACE/AII	Change taste	Stop Bang
Urea	IgMIgD-	B-Blockers	Change in HrQOL	Insomnia (ISI)
creatinemia	IgMIgD+	Calcium channel	Fatigue (Chalder)	SF-physical score
Acido urico	CD27-	Potassium sparing	Chalder Total Points	SF-12 mental score
Calcio	IgMIgD-	Thiazide drug	Fatigue	Record time
Fosforo	IgMIgD+	Metformin	Heart rate	RDI
LDH	Naive	Insulin	SB P	RDI > 5
CK total	Tran	Hipolipemic drug	DBP	RDI > 15
Col Tot	Linfo Count	Z drugs	SpO2	RDI>30
Col HDL	B cell Count	IRSS	PEF	T90
Col LDL	IL12	Weight	Abnormal Handgrip	ODI
Trigliceridos	TNF	Height	BMI	mean O2
Bili Tot	IL6	BMI	Neck circumf.	Lowest O2
Bili dir	IL1b	Neck circumf.	Waist circumf	Min pulse
Bili indir	IL-8	Waist circumf	Abnormal waist	Max pulse
Prot Tot	C3a	Hip circumf.	Hip circumf.	Ave. pulse
Albumina	C4a	preEuroQoL	SBP	RDI > 5
Globulina	C5a	Fever	SBP/DBP	TNS
rel A/G	Severity	Headache	SBP>140	SD TNS
GOT (AST)	ARDS	Chest pain	DBP	TST
GPT (ALT)	Evaluation date	Sore throat	DBP>90	SD TST
GGT	Weeks since Dx	Cough	Glucose	Onset Latency
Fosfatasa Alcalina	IgM SARS-COV-2	Dysnea	Abnormal glucose	SD onset
Insulina	IgG SARS-COV-2	Polypnea	Total cholesterol	Sleep Efficiency
HOMA	Age	Mialgia	HDL-cholesterol	SD Sleep Eff
Hematocrito	Gender	Desaturation	Abnormal HDL	WASO
Hemoglobina	Smoker	Abdominal pain	LDL-Cholesterol	SD WASO
Eritrocitos	Pack/year	Dyarrhea	Trygliceridemia	Arousal
VCM	Alcohol	Change smell	Abnormal TG	SD arousal
HCM	Scholar	Change taste	Total hiperlipemia	CFI

ID				
CHCM	Rural area	postEuroQoL	Gender	M10
Leucocitos	HTA	Fever	sMet (point)	L5
%Linfocitos	Insulin Resistance	Headache	Smet (yes/no)	RA
%Monocitos	T2DM	Chest pain	mMRC	Interdialy stability
%Granulocitos	Heart Failure	Sore throat	SATED	Intradaily stability
Plaquetas	COPD	Cough	Pittsburg	Mesor
Linfo Count	Cancer	Dysnea	ESS	Amplitude
Monocitos	CKD	Polypnea	SQALI	Acrophase
Granulocitos	Afib	Mialgia	MEQ	Time
IgM SARS-COV-2	Stroke	Desaturation	MEQ (cronotipe)	
IgG SARS-COV-2	CHD	Abdominal pain	HADS-A	

4. Características por subset de datos

***Tabla Anexo III**. Características por subset de datos*

df_actigrafía	df_clinica	df_cuestionario	df_espirometria	df_demografia	df_hemograma	df_poligrafia	df_sintomatología	df_subcell	df_tac
TNS	Glicemia	mMRC	preFVC (L) real	Age	Hematocrito	Record time	preEuroQoL	B cells	Chest CT
SD TNS	Urea	SATED	preFEV1 (L) real	Gender	Hemoglobina	RDI	Fever	CD27+	Abnormal Chest CT
TST	creatinemia	Pittsburg	preVEF1/FVC real (%)	Smoker	Eritrocitos	RDI > 5	Headache	IgMIgD-	Grond-glass
SD TST	Acido urico	ESS	preFEF 25-75% (L/s)real	Pack/year	VCM	RDI > 15	Chest pain	IgMIgD+	Mixed ground-glass
Onset Latency	Calcio	SQALI	preFEF max (L/s) real	Alcohol	HCM	RDI>30	Sore throat	CD27-	Consolidation
SD onset	Fosforo	MEQ	preFVC (L) % Teo	Scholar	CHCM	T90	Cough	IgMIgD-.1	Interlobular thickening
Sleep Efficiency	LDH	MEQ (cronotipe)	preFEV1 (L) % Teo	Rural area	Leucocitos	ODI	Dysnea	IgMIgD+.1	Bronchiectasis
SD Sleep Eff	CK total	HADS-A	preVEF1/FVC % Teo	HTA	%Linfocitos	mean O2	Polypnea	Naive	Atelectasis
WASO	Col Tot	HADS-D	preFEF 25-75% % Teo	Insulin Resistance	%Monocitos	Lowest O2	Mialgia	Tran	Solid nodule
SD WASO	Col HDL	Beck Depression	preFEF max (L/s) % Teo	T2DM	%Granulocitos	Min pulse	Desaturation		Nonsolid nodule
Arousal	Col LDL	Stop Bang	postFVC (L) real	Heart Failure	Plaquetas	Max pulse	Abdominal pain		Number of Lobes affected
SD arousal	Trigliceridos	Insomnia (ISI)	postFEV1 (L) real	COPD	Linfo Count	Ave. pulse	Dyarrhea		Reticular lesions
CFI	Bili Tot	SF-physical score	postVEF1/FVC real (%)	Cancer	Monocitos		Change smell		Fibrotic lesions
M10	Bili dir	SF-12 mental score	postFEF 25-75% (L/s)real	CKD	Granulocitos		Change taste		Air trapping
L5	Bili indir		postFEF max (L/s) real	Afib					none
RA	Prot Tot		postFVC (L) (%)	Stroke					TSS Total
Interdialy stability	Albumina		postFEV1 (L) % Teo	CHD					

df_actigrafía	df_clinica	df_cuestionario	df_espirometria	df_demografia	df_hemograma	df_poligrafia	df_sintomatología	df_subcell	df_tac
Intradaily stability	Globulina		postVEF1/FVC% Teo	NAFLD					
Mesor	rel A/G		postFEF 25-75% (L/s)% Teo	Hypothyroidism					
Amplitude	GOT (AST)		postFEF max (L/s) % Teo	ACE/AII					
Acrophase	GPT (ALT)		FEV1<70%	B-Blockers					
	GGT		DLCO<80	Calcium channel					
	Fosfatasa Alcalina		6MWT	Potassium sparing					
	Insulina		HR (basal)	Thiazide drug					
	HOMA		HR (final)	Metformin					
			Spo2 (basal)	Insulin					
			Spo2 (final)	Hipolipemic drug					
			Drop in Spo2	Z drugs					
			Drop >3%	IRSS					
			Borg basal	Weight					
			Borg final	Height					
			Fatigue basal	BMI					
			Fatigue final	Neck circumf.					
				Waist circumf					
				Hip circumf.					