UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCUELA TÉCNICA SUPERIOR INGENIERÍA INDUSTRIAL VALENCIA

TRABAJO FIN DE GRADO EN INGENIERÍA BIOMÉDICA

DESARROLLO DE UN ALGORITMO DE APRENDIZAJE AUTOMÁTICO PARA PREDICCIÓN DE EVOLUCIÓN DE PACIENTES HOSPITALIZADOS POR COVID-19

AUTORA: BELÉN BALLESTER VICIANO

TUTORA: VALERY NARANJO ORNEDO

COTUTOR: JORGE IGUAL GARCÍA

Curso Académico: 2020-21

AGRADECIMIENTOS

A mi apoyo más grande e incondicional, por enseñarme que lo imposible solo es un poquito más difícil. Mamá, papá, Guille, si hoy soy ingeniera biomédica es gracias a vosotros.

A todas las personas que me han acompañado en este camino y no me han dejado dudar de mí. Vosotros y vosotras sabéis quienes sois… es una suerte enorme teneros en mi vida.

A mis tutores, por brindarme la oportunidad de trabajar en un proyecto tan cercano a la realidad clínica, por la ayuda y por la paciencia.

RESUMEN

En el presente proyecto de final de grado se pretende desarrollar un algoritmo capaz de etiquetar según su gravedad a pacientes hospitalizados por COVID-19. Este algoritmo busca abrirse paso en situaciones de colapso hospitalario, ofreciendo una ayuda diagnóstica intuitiva al personal clínico para tratar y monitorizar a dichos pacientes.

Para el desarrollo del trabajo se ha tomado la base de datos del proyecto "COVID DATA SAVE LIVES", proporcionada a la comunidad científica por el grupo hospitalario HM Hospitales en abril de 2020. Se ha realizado un análisis exploratorio sobre dichos datos y se ha tratado de encontrar en ellos aquellas variables que puedan ser discriminantes a la hora de predecir cómo evolucionará un paciente hospitalizado por COVID-19.

Durante el análisis exploratorio de los datos se encontraron diversas limitaciones que podrían afectar al correcto desarrollo del predictor; sin embargo, se ha logrado finalmente plantear un algoritmo de predicción capaz de etiquetar con éxito a un 74,8% de los pacientes enfermos.

Palabras clave: COVID-19, SARS-CoV-2, Inteligencia Artificial, Macrodatos, Predicción de enfermedades con algoritmos de Aprendizaje Automático, Clasificadores, Análisis exploratorio de datos

RESUM

En el present projecte de final de grau es pretén desenvolupar un algoritme capaç de etiquetar segons la seua gravetat a pacients hospitalitzats per COVID-19. Aquest algoritme busca obrir-se pas en situacions de col·lapse hospitalari, oferint una ajuda diagnòstica intuïtiva al personal clínic per a un millor tractament i monitorització dels pacients.

Per al desenvolupament del treball s'ha pres la base de dades del projecte "COVID DATA SAVE LIVES", proporcionada a la comunitat científica pel grup hospitalari HM Hospitales a l'abril de 2020. S'ha realitzat un anàlisi exploratori d'aqueixes dades i s'ha tractat de trobar en elles les variables que pogueren ser discriminants a l'hora de predir com evolucionarà un pacient hospitalitzat per COVID-19.

Durant l'anàlisi exploratori de les dades es van trobar diverses limitacions que podrien haver afectat al correcte desenvolupament del predictor; no obstant això, s'ha aconseguit finalment plantejar un algoritme capaç d'etiquetar amb èxit a un 74,8% dels pacients malalts.

Paraules clau: COVID-19, SARS-CoV-2, Intel·ligència Artificial, Macrodades, Predicció de malalties amb algoritmes de Aprenentatge Automàtic, Classificadors, Anàlisi exploratori de dades

ABSTRACT

The aim of this final degree project is to develop an algorithm capable of labelling patients hospitalized for COVID-19 according to their severity. This algorithm seeks to break through in hospital collapse situations, offering an intuitive diagnostic aid to clinical staff to treat and monitor these patients.

For the development of the project, the database of the "COVID DATA SAVE LIVES" project provided to the scientific community by the HM Hospitales hospital group in April 2020 was taken. An exploratory analysis of the data was carried out and an attempt was made to find in them variables that may be discriminating when predicting how a patient hospitalized for COVID-19 will evolve.

During the exploratory analysis of the data, different limitations that could affect the correct development of the predictor were found; however, it was finally possible to develop a prediction algorithm with a 74,8% of accuracy.

Keywords: COVID-19, SARS-CoV-2, Artificial Intelligence, Big Data, Disease prediction with Machine Learning algorithms, Classifiers, Exploratory data analysis

ACRÓNIMOS

SARS-CoV-2	*Severe acute Respiratory Syndrome Coronavirus 2*
COVID-19	Enfermedad por coronavirus 2019
IA	Inteligencia Artificial
ML	*Machine Learning*
ID	Identificador
FC	Frecuencia Cardíaca
PA	Presión Arterial
PaO_2	Presión arterial parcial de Oxígeno
OMS	Organización Mundial de la Salud
UCI	Unidad de Cuidados Intensivos
HCE	Historia Clínica Electrónica

ÍNDICE GENERAL

II. Presupuesto

III. Anexos

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

MEMORIA

Desarrollo de un algoritmo de aprendizaje automático para predicción de evolución de pacientes hospitalizados por COVID-19

Documento I

Belén Ballester Viciano

Grado en Ingeniería Biomédica

Curso académico 2020 - 2021

1. OBJETIVOS

El objetivo principal de este proyecto es el de desarrollar un modelo capaz de predecir cómo evolucionará un paciente con COVID-19, utilizando algoritmos propios de la inteligencia artificial y técnicas de *big data*. El propósito es que dicho sistema sea capaz de clasificar a los individuos sobre los que se le proporcionen una serie de datos con la mayor precisión posible, utilizando algoritmos supervisados de *machine learning*.

Para llegar a dicho fin, se tomará la base de datos proporcionada por HM Hospitales y se llevará a cabo un proceso exhaustivo de comprensión y limpieza de los datos, construyéndose un modelo reducido con información de calidad. Tras ello, se procederá a analizar los datos del conjunto reducido de datos y se extraerá de este la información pertinente para poder desarrollar un algoritmo de predicción fiable. De este modo, se incluye entre los objetivos también la realización de un correcto procedimiento de análisis exploratorio de los datos.

Se busca con todo esto descubrir patrones o correlaciones entre variables que puedan ser clave para estratificar el riesgo de un paciente con COVID-19, de forma que su tratamiento sea lo más eficiente posible.

Por otra parte, a modo de asegurar el desarrollo de un algoritmo que se adapte de la mejor forma posible a dichos datos, se configurarán y compararán distintos tipos de algoritmos.

2. INTRODUCCIÓN

A fin de comprender con claridad cada una de las partes del presente trabajo, se pretende en este capítulo realizar un breve resumen del contexto en que se enmarca, de sus antecedentes y de distintos aspectos que resultarán clave para su entendimiento.

Índice de contenidos

2.1. MOTIVACIÓN

El tratamiento de los pacientes por parte de los servicios sanitarios es un proceso complejo y que demanda la mayor calidad posible. Desde que se detecta una anomalía en el estado de salud del paciente, pasando por el diagnóstico hasta su tratamiento y ulterior recuperación, todos los procesos deben monitorizarse con especial cuidado. Sin embargo, esto trae consigo un enorme consumo de recursos que en ocasiones los centros sanitarios no pueden asumir.

La pandemia de COVID-19 llegó a hospitales de todo el mundo en un momento en que no existía previsión ni medios para tratarla. Los sistemas sanitarios se encontraron con falta de respiradores, de pruebas diagnósticas, de camas y de equipos de protección individuales (EPIs) para proteger a los trabajadores sanitarios. La falta de recursos fue pronto evidente y la saturación en hospitales, clínicas, residencias y otros centros de atención sanitaria fue creciendo.

Ante esta situación, los profesionales sanitarios se vieron obligados a tomar decisiones difíciles, complejas y sobre las cuales no tenían una total certeza. De hecho, se conocen testimonios de profesionales sanitarios que han tenido que decidir una reasignación de recursos sin un criterio totalmente establecido o detener el auxilio respiratorio de pacientes graves para poder tratar a pacientes con más posibilidades de sobrevivir.

Una correcta previsión y gestión podrían haber solventado parte del problema, pero ante situaciones imprevistas como la que aconteció al inicio de la pandemia de COVID-19 esto resulta excesivamente difícil. Sin embargo, un apoyo significativo podría haber sido un sistema de ayuda a la toma de decisiones capaz de evaluar a cada paciente. De este modo, los profesionales sanitarios podrían haber tomado decisiones más rápidas y basadas en todas las evidencias conocidas para decidir acerca de la hospitalización de un paciente, de si era seguro darle el alta o si era recomendable hacer un seguimiento del tratamiento desde su propio hogar.

Una toma de decisiones eficaz es crucial en el proceso asistencial, pues ayuda a una mayor descongestión de centros sanitarios en situaciones de colapso, permite un tratamiento de calidad y acorde a las necesidades de los pacientes y ayuda a reducir el consumo de recursos y capital.

Una buena forma de obtener buenos resultados en la toma de decisiones sin sucumbir al error humano es recurrir a algoritmos de inteligencia artificial (IA). En la actualidad existen múltiples algoritmos de este tipo en el ámbito de la salud, y en el caso particular de la toma de decisiones se ha observado que ofrecen un enorme potencial para llevar a cabo dicha tarea de forma automatizada. De hecho, autores afirman que, con unos buenos datos a su entrada, estos algoritmos tienen un poder de diagnóstico de enfermedades y de selección de tratamientos excelente [1].

En cuanto respecta al uso de estos algoritmos en evaluación de COVID-19, la falta de conocimiento inicial del virus dificultó el proceso, pero en la actualidad se tienen bases de datos amplias que permiten generar algoritmos sólidos. Estos algoritmos permitirían conocer más la enfermedad y contribuirían a mejorar la toma de decisiones clínica en contextos como los descritos anteriormente.

Así pues, existen en la actualidad bases de datos abiertas a la comunidad científica para estudiar distintos aspectos de esta enfermedad, los cuales con ayuda de la IA pueden significar una ayuda inestimable en el avance de la actual pandemia. De este modo, el tema y motivación del presente trabajo surge en busca de plantear un algoritmo de IA capaz de ayudar en la predicción y toma de decisiones de pacientes relativa a enfermos por COVID-19.

2.2. DESCRIPCIÓN DEL PROBLEMA

La situación de pandemia que se iniciaba en marzo de 2020 a causa de la enfermedad por coronavirus (COVID-19) ha tenido un enorme impacto en los sistemas sanitarios de todo el mundo. Esta enfermedad, completamente nueva en aquel momento, generó una gran carga de trabajo que saturó hospitales y centros sanitarios. Mientras tanto, el desconocimiento, la desinformación y el miedo se abrían paso en todos los sectores de la sociedad.

El total desconocimiento de la virulencia y alcance de esta enfermedad exigió a la comunidad científica mundial un esfuerzo conjunto para determinar la epidemiología, características microbiológicas y clínicas de la COVID-19. Los hospitales y centros de salud contribuyeron, además de con toda la labor asistencial y cuidado médico, con valiosos datos e información que permitirían avanzar en el conocimiento clínico de la enfermedad y en el manejo de los pacientes infectados por ella.

2.2.1. Pandemia de COVID-19.

El 31 de diciembre de 2019, China comunicaba a la agencia de la Organización de las Naciones Unidas (ONU) la existencia de varios casos de una neumonía viral de origen desconocido en la ciudad de Wuhan. Durante los siguientes días se identificaba el agente etiológico como un nuevo coronavirus, el 2019-nCoV (*2019-novel coronavirus*), al que posteriormente se le denominaría SARS-CoV-2 (*Severe Acute Respiratory Syndrome Coronavirus* 2). A la enfermedad que produce se la denominó COVID-19 (C*oronavirus disease 2019*) [2], [3].

Rápidamente comenzaron a multiplicarse los casos en la República Popular China, notificándose los primeros casos confirmados fuera del país el 13 y 16 de enero. Ante esto, la Organización Mundial de la Salud (OMS) publicó la primera alerta epidemiológica de lo que sería la pandemia de COVID-19 [4].

A pesar de los esfuerzos llevados a cabo por frenar el avance de la enfermedad, hacia finales de enero se reportaban las primeras muertes y el número de contagiados aumentaba exponencialmente. De este modo, el día 30 de enero la OMS declaraba la COVID-19 como una emergencia de salud pública de importancia internacional al haberse notificado casos en todas las regiones de la OMS en tan solo un mes [3], [4].

El SARS-CoV-2 se extendió por todas las regiones del mundo, incrementándose bruscamente el número de infectados y muertos. A estos efectos, el 11 de marzo la OMS caracterizaba la enfermedad como pandemia y el 13 de marzo Europa era declarada el epicentro de la pandemia [3], [4]. Seguidamente, países de todo el mundo decretaron el confinamiento domiciliario obligatorio, la limitación de la libertad de circulación y el cese de las actividades no esenciales. En España se establecían estas medidas a través del estado de alarma que se decretó el 14 de marzo.

A nivel mundial comenzó una crisis sanitaria en que la presión asistencial, la saturación permanente en atención primaria, la creciente falta de recursos y de camas en las Unidades de Cuidados Intensivos (UCI), así como el desconocimiento de la enfermedad para tratar adecuadamente a los pacientes, llevó al colapso de muchos sistemas sanitarios.

Sin embargo, gracias a las medidas adoptadas por los gobiernos y al trabajo de los sanitarios e investigadores científicos, se logró estabilizar la curva de contagios y ralentizar el avance de la enfermedad. El número de contagios y muertes descendió significativamente en toda Europa, relajándose las medidas a partir del mes de mayo en la mayoría de los países del continente. En España, el estado de alarma terminaba el día 21 de junio.

Mientras tanto, grupos de investigadores y farmacéuticas de todo el mundo continuaban trabajando por desarrollar una vacuna contra la COVID-19, así como técnicas diagnósticas más rápidas y tratamientos contra el virus. La primera vacuna que se autorizaría en Europa sería la de BioNTech-Pfizer, tras obtener la aprobación de la Agencia Europea del Medicamento (EMA) el 21 de diciembre de 2020 [5].

Tras la primera ola de contagios se sucedieron en distintos países hasta 3 olas más, siendo la tercera ola en España tras las vacaciones de Navidad, tal y como se aprecia en la Figura 1. Sin embargo, aunque el número de nuevos casos y fallecimientos en esta ola fue muy elevado, la campaña de vacunación en España comenzó el 27 de diciembre, lo que supuso una enorme ventaja respecto a las anteriores olas de la pandemia.

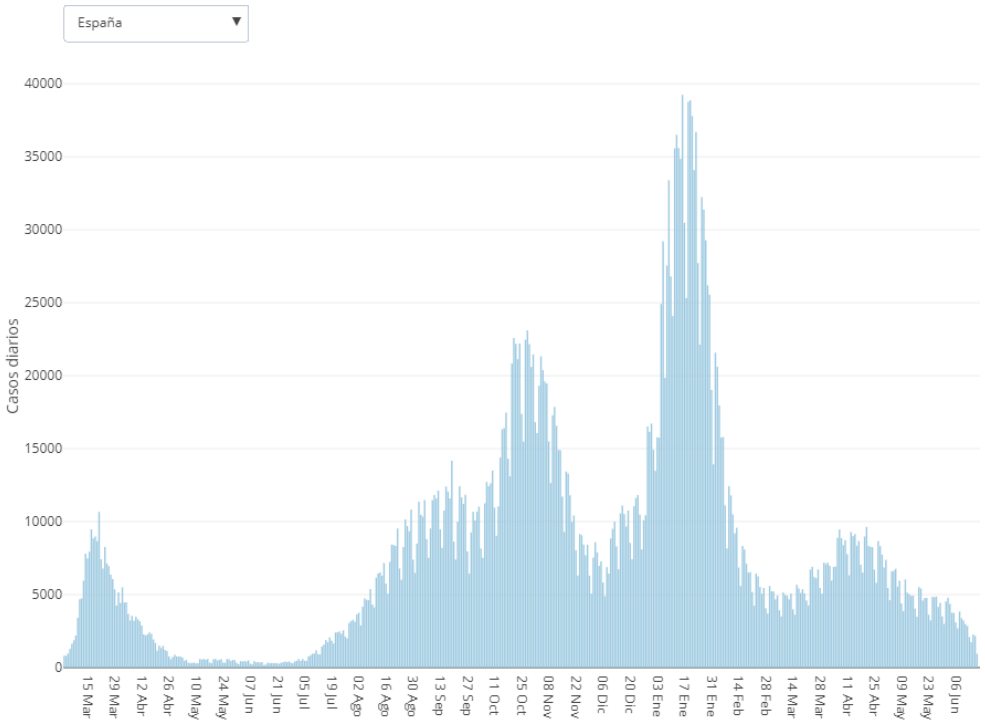

Figura 1. *Curva epidémica de la pandemia en España. Datos obtenidos a partir de datos individualizados notificados a la Red Nacional de Vigilancia Epidemiológica (RENAVE) a fecha 20 de junio de 2021. Figura tomada a través de la aplicación informática SiViEs* [6].

Para el 18 de junio de 2021, según el Informe de actividad del proceso de vacunación del Ministerio de Sanidad, en España se estaban utilizando 4 tipos de vacuna: Pfizer-BioNTech, Moderna, AstraZeneca-Oxford y Janssen. Un 47,8% de la población había recibido al menos una dosis de la vacuna y un 29,4% la pauta completa [7].

A nivel mundial y para estas mismas fechas, se habían registrado desde el inicio de la pandemia más de 178 millones de casos confirmados, siendo el número de fallecidos a causa de esta enfermedad superior a 3,86 millones de personas [8].

2.2.2. SARS-CoV-2

El coronavirus SARS-CoV-2 es el patógeno que causa la enfermedad COVID-19 y que ha provocado la pandemia mundial más grande de la historia reciente. Fue aislado por primera vez a principios de 2020 en el fluido resultante del lavado broncoalveolar de los primeros pacientes con síntomas de COVID-19 en Wuhan [3].

Los coronavirus son un grupo de virus de la familia *Coronaviridae*, del orden de los *Nidovirales.* Esta familia se subdivide en 4 diferentes géneros, de los cuales solamente los alfa-coronavirus y los beta-coronavirus tienen capacidad para infectar a humanos [9].

El SARS-CoV-2 por su parte es un tipo de coronavirus que se enmarca dentro de los beta-coronavirus. Esto presupone un origen zoonótico, es decir, por transmisión entre un huésped animal y un humano; y a favor de esta hipótesis se conocen distintos coronavirus de murciélago muestreados en la provincia de Yunnan (China), que poseen estrecha relación con el SARS-CoV-2, siendo su secuencia genómica idéntica en varios casos superior al 95% [9], [10]. A pesar de ello existe controversia acerca de su origen y múltiples fuentes defienden que pueda tener su origen en un laboratorio. En la actualidad, el conocimiento sobre el origen animal del virus sigue siendo incompleto.

Según lo que se ha mencionado, este patógeno sigue la estructura general de los coronavirus. Se trata de virus esféricos o pleomorfos, de entre 80 y 120 nm y cuyo genoma se compone por una sola hebra de ácido ribonucleico (ARN) sencilla no segmentada y de polaridad positiva con un tamaño de unos 30.000 nucleótidos. Esta hebra se encuentra rodeada de una membrana compuesta de distintos lípidos y glicoproteínas que le confieren un aspecto similar al de una corona, fácilmente perceptible a través de microscopía electrónica (véase la Figura 2) [3], [11].

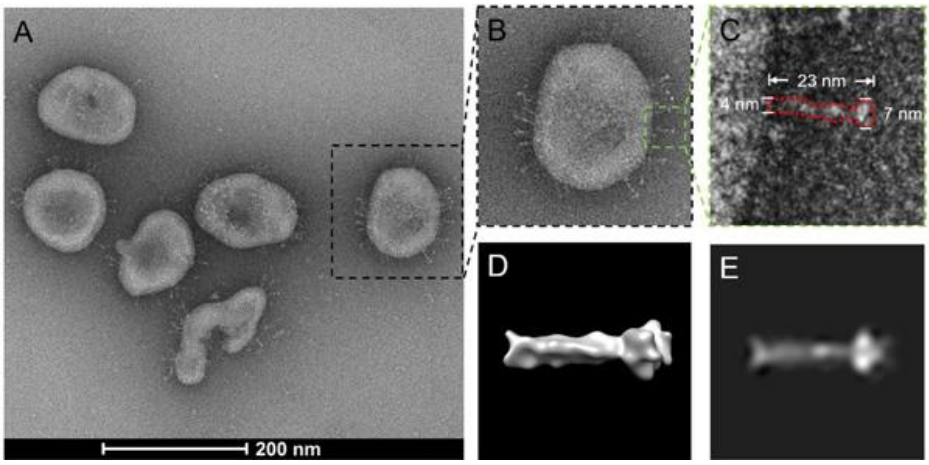

Figura 2. *Imagen del SARS-CoV-2 obtenida en microscopio electrónico. A) Virus de SARS-CoV-2 aislados. B) Ampliación de un coronavirus. Se aprecian las espículas virales alrededor de su envoltura. C – E) Caracterización de las espículas formadas por las glicoproteínas del SARS-CoV-2. Figura obtenida a partir del periódico online Redacción Médica* [12].

Respecto a la transmisión del SARS-CoV-2, este accede al cuerpo humano a través de las vías respiratorias, por exposición de un individuo sano a objetos contaminados o a individuos infectados, ya sean sintomáticos o no. La transmisión se produce primordialmente por el contacto directo o indirecto de las llamadas gotitas de Flügge, expelidas cuando una persona infectada habla, tose, estornuda o simplemente respira [3].

Cuando el virus se introduce en el organismo, accede inicialmente a las células epiteliales de las vías respiratorias superiores e inferiores, fijándose a ellas a través de las glicoproteínas de su superficie. Se produce entonces una fusión con la membrana celular de la célula huésped y el virus libera en el interior de esta su ARN, a fin de que se replique y se sintetice en forma de proteínas. La maquinaria celular humana confunde entonces este ARN viral con el ARN de la propia célula y comienza a formar virus idénticos en el interior de la célula infectada, los cuales destruyen la célula a continuación y salen al medio extracelular para infectar nuevas células [3].

Se ha observado que esta infección celular se produce a velocidad exponencial y que se extiende rápidamente a distintos órganos y sistemas del cuerpo más allá de las vías respiratorias. Sin embargo, se ha observado que esto se produce de forma selectiva, ya que cuando se trata de casos graves parece que el daño se extiende solamente hacia algunas zonas como el corazón, el hígado, los riñones y ciertas partes del sistema neurológico [13].

Las características clínicas de la infección por SARS-CoV-2 comprenden desde un cuadro asintomático hasta una neumonía severa que produzca insuficiencia respiratoria a causa del síndrome de distrés respiratorio del adulto (SDRA) y disfunción multiorgánica. Al menos 5 días son necesarios para que aparezcan los primeros síntomas [3], [9]. La mayor parte de las manifestaciones clínicas que se han reportado para la enfermedad de la COVID-19 quedan reflejadas en la Tabla 1.

Es importante mencionar también las distintas formas de diagnóstico de COVID-19 utilizadas actualmente. En orden de mayor a menor fiabilidad de detección inmediata, las pruebas son reacción en cadena de la polimerasa (PCR), prueba serológica de antígenos y prueba serológica de anticuerpos.

La COVID-19 afecta de diversas formas a cada persona, dependiendo de muchos factores el que aparezca o no un síntoma o la intensidad de este. Según el Informe del Grupo de Análisis Científico de Coronavirus del Instituto de Salud Carlos III (GACC-ISCIII) del 1 de junio de 2020, generalmente las personas mayores sufren más la enfermedad y muestran una peor evolución, mientras que también afecta en mayor medida a hombres [14]. Además, existen distintos factores de riesgo que agravan la enfermedad, como la presencia de enfermedades crónicas y ciertas comorbilidades.

En este informe se establecieron como factores de riesgo las enfermedades cardiovasculares (cardiopatías, hipertensión arterial…), la diabetes, las enfermedades respiratorias crónicas (enfermedad pulmonar obstructiva crónica (EPOC), asma…), las enfermedades renales, cánceres, la inmunosupresión (pacientes oncológicos, trasplantados…), enfermedades neurológicas, el sobrepeso o la obesidad y el tabaquismo. Más recientemente se han encontrado evidencias de que son también factores de riesgo el estar embarazada, los trastornos por uso de sustancias y ser seropositivo (infectado por VIH) [15].

Tabla 1. *Manifestaciones clínicas reportadas en pacientes con COVID-19. Tabla elaborada a partir de* [3], [16], [17].

Manifestaciones clínicas de la COVID-19	***Más habituales***	**Fiebre**
		Tos seca
		Fatiga, Astenia
		Biomarcadores: linfopenia; aumento leve de LDH, del dímero D y del tiempo de protrombina
	Menos frecuentes	Dolores musculares y articulares
		Odinofagia
		Cefalea
		Pérdida del olfato o del gusto
		Escalofríos o vértigo
		Erupciones cutáneas, pérdida del color en los dedos de manos y/o pies
		Diarrea
		Náuseas o vómitos
		Congestión nasal
		Hemoptisis
		Conjuntivitis
		Radiografía anormal
		Biomarcadores: disminución de saturación de oxígeno en sangre e hipoxia; aumento de dímero D, LDH y transaminasas
	Graves	Disnea
		Dolor o presión persistente en el pecho
		Incapacidad para hablar o moverse
		Confusión
		Pérdida de apetito
		SDRA, sepsis, insuficiencia renal y/o cardíaca aguda
		Biomarcadores: aumento de la ferritina, de dímero D, PCReactiva, troponina, creatinina, tiempo de coagulación e interleucina 6

2.3. INTELIGENCIA ARTIFICIAL

En un mundo donde cada vez coge más peso la digitalización de los datos y la automatización, la inteligencia artificial (IA) adquiere un papel importante. La IA es una forma de hacer que una máquina se comporte y actúe de forma inteligente; es una ciencia que busca teorías y metodologías que permitan a las máquinas entender el mundo y, en consecuencia, reaccionar a las situaciones de la misma forma en que lo haría un humano [18]. De acuerdo con esto y dependiendo del entorno en que se encuentre, una máquina inteligente podrá mostrar sus habilidades a través de distintos medios como el razonamiento, la deducción, el aprendizaje, la creatividad o la planificación [19].

Los sistemas que utilizan IA son además capaces de adaptar su comportamiento, de manera que cuando reciben nuevos datos o nueva información, son capaces de procesarlos e inferir una respuesta en función de respuestas anteriores, trabajando de manera autónoma y corrigiendo errores previos.

En el campo de la salud, la IA ofrece aplicaciones prometedoras entre las que destacan la ayuda a la toma de decisiones, la detección de anomalías, la detección de biomarcadores, la atención médica en áreas remotas, la cirugía robótica, la descongestión de centros sanitarios, etc. [1], [20] En este mismo ámbito, la IA permite analizar grandes cantidades de datos y buscar patrones que puedan suponer un avance para el diagnóstico, el tratamiento y la predicción de resultados en distintos escenarios clínicos.

La disponibilidad de paquetes de aprendizaje automático o *machine learning* (ML) de código abierto y el acceso libre y gratuito a datos clínicos para la comunidad científica son factores clave para el éxito de estas aplicaciones, además de ser uno de los factores que ha impulsado el crecimiento de la IA [1].

2.3.1. Aprendizaje Automático

El aprendizaje automático o ML es la forma más popular de IA [18]. Se entiende como un método de análisis de datos que automatiza la construcción de modelos analíticos y cuyo objetivo principal consiste en desarrollar técnicas que permitan que las máquinas o sistemas aprendan a partir de datos, identifiquen patrones y hagan predicciones acerca de ellos con la mínima intervención humana posible [19], [21]. Un problema de este tipo de modelos es que se encuentran limitados al valor de los datos utilizados [18]. En la Figura 3 se aprecia el desarrollo general de un sistema de ML.

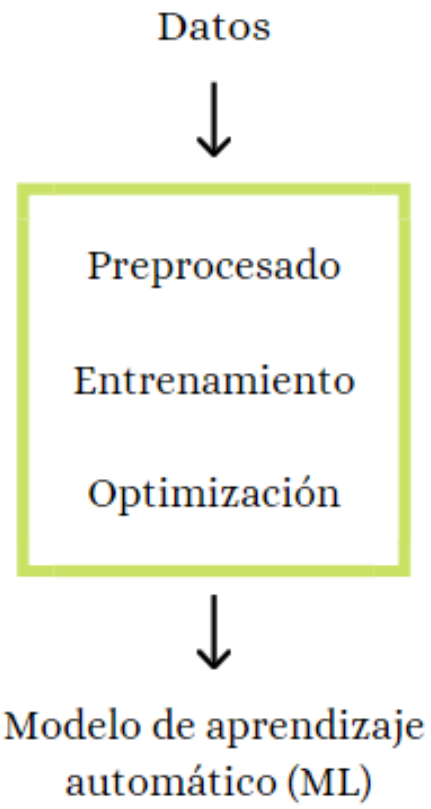

***Figura 3**. Sistema típico de desarrollo de un modelo de ML. Figura obtenida por elaboración propia.*

Es importante mencionar que existen distintos métodos de ML clasificados en función de la naturaleza de los datos empleados para el aprendizaje o de la retroalimentación disponible. Cada método se enfoca en tratar un problema distinto. Entre ellos destacan:

- **Aprendizaje supervisado**

A partir de ejemplos etiquetados (conjunto de datos de entrada con los resultados correctos correspondientes) un algoritmo aprende a asignar a cada nueva entrada el resultado deseado.

El algoritmo aprende a base de comparar el resultado que ha obtenido con el resultado real proporcionado por la etiqueta y modifica el modelo en consecuencia. Cuando el modelo queda definido, utiliza patrones para predecir los valores de la etiqueta de datos nuevos que están sin etiquetar. Cuanto mayor es el conjunto de datos, mejor es capaz el modelo de generalizar.

- **Aprendizaje no supervisado**

Al algoritmo no se le proporcionan datos etiquetados a su entrada, de modo que tiene que descubrir por sí mismo patrones en los datos que le permitan diferenciarlos. El objetivo de este método consiste en explorar los datos y encontrar alguna estructura en su interior.

Estos algoritmos se utilizan fundamentalmente para identificar objetos y patrones dentro del conjunto de datos que se les ofrece, en lugar de para predecir.

- **Aprendizaje semisupervisado**

Se combinan el aprendizaje supervisado y el no supervisado: al algoritmo se le proporcionan algunos datos que están etiquetados y otros que no lo están. Generalmente la mayoría de los datos no están etiquetados, pues con este método se pretende reducir el coste de computación de los métodos de aprendizaje supervisado.

- **Aprendizaje por refuerzo**

Este método se centra en que el algoritmo aprenda por medio de la interacción dinámica con el entorno, a través de ensayo y error. La retroalimentación es la que da respuestas al sistema, que deberá de encargarse de optimizar la toma de decisiones de modo que se consiga la máxima recompensa en sus respuestas.

2.3.1.1. Aprendizaje Supervisado y Métricas de Evaluación

El aprendizaje no supervisado tiene aplicaciones en medicina como la detección de anomalías o la reducción dimensional de los datos [1]. Sin embargo, no permite a los algoritmos aprender de sus aciertos o errores, pues el valor a la salida no puede compararse con un valor de referencia (etiqueta).

Por su parte, el aprendizaje supervisado permite un aprendizaje a través de los errores, lo que lo hace especialmente atractivo para aplicaciones de regresión y clasificación de datos [1]. Por una parte, la regresión permite relacionar variables, predecir su evolución e inferir resultados futuros basándose en los resultados disponibles. Por otra parte, la clasificación de datos utiliza el resultado de dichas predicciones para determinar la clase de una muestra y agruparla con muestras similares.

Cuando se trabaja con algoritmos supervisados, es necesario que existan unas métricas de evaluación que den estimaciones objetivas sobre cómo de bueno o malo es un clasificador. Estas métricas devuelven el grado de precisión, fiabilidad y sensibilidad del algoritmo seleccionado, y se basan en la comparación entre los resultados obtenidos y los que se esperaban (los contenidos en las etiquetas):

- Verdadero Positivo (TP): el resultado predicho y el esperado fueron positivos.
- Verdadero Negativo (TN): el resultado predicho y el esperado fueron negativos.
- Falso Positivo (FP): el resultado predicho fue positivo, pero el esperado era negativo.
- Falso Negativo (FN): el resultado predicho fue negativo, pero el esperado era positivo.

Si estos valores se ordenan en una tabla de contingencia, es fácil observar cuándo el clasificador está actuando correctamente. En la Figura 4 se puede observar un ejemplo de esta matriz de contingencia aplicada al ámbito de la medicina. Sin embargo, cabe mencionar que esto es útil solamente para algoritmos de clasificación binaria; si se trata con clasificadores más complejos, la herramienta que se deberá emplear es la matriz de confusión, que permite visualizar el desempeño del algoritmo seleccionado (véase la Figura 5).

Tipos de diagnósticos		Enfermedad *(realidad)*	
		Ausente	Presente
Prueba diagnóstica *(predicción)*	Negativa	TN	FN
	Positiva	FP	TP

***Figura 4**. Tabla de contingencia que contrasta la presencia de una enfermedad (valor esperado) con el resultado de una prueba diagnóstica (valor predicho) utilizando valores de métricas de evaluación. Figura obtenida por elaboración propia.*

		Valor predicho		
		Gato	Perro	Conejo
Valor real	Gato	8	2	0
	Perro	4	5	1
	Conejo	0	1	10

***Figura 5.** Ejemplo de matriz de confusión. Figura obtenida por elaboración propia.*

El verdadero fin de estos valores radica en su uso para conocer en detalle la exactitud de los sistemas de clasificación. En el ámbito clínico, la toma de decisiones es un aspecto extremadamente complejo, por lo que conocer estos marcadores es clave para conocer la capacidad del sistema para clasificar patrones en distintas categorías o estados [22].

Así pues, con base a estos valores se puede conocer la probabilidad de hacer una predicción correcta. Se establecen de este modo las siguientes métricas de evaluación:

- **Precisión:** indica la proporción de aciertos positivos sobre el total de positivos predichos.

$$Precisión = \frac{TP}{TP + FP} \quad (1)$$

- **Exactitud:** indica la proporción de predicciones correctas.

$$Exactitud = \frac{TP + TN}{Total} \quad (2)$$

- **Especificidad:** indica la proporción de aciertos negativos sobre el total de negativos predichos.

$$Especificidad = \frac{TN}{TN + FP} \tag{3}$$

- **Sensibilidad** (fracción de TP)**:** indica la proporción de aciertos positivos sobre el total de positivos.

$$Sensibilidad = \frac{TP}{TP + FN} \tag{4}$$

Recibe especial importancia la curva ROC, que representa la fracción de verdaderos positivos (sensibilidad) frente a la fracción de falsos positivos (1-especificidad). Cuando se obtiene un clasificador, se desea tener las mayores sensibilidades y especificidades posibles. Si se observa la Figura 6, resulta evidente que el modelo A dará mejores resultados que el modelo B.

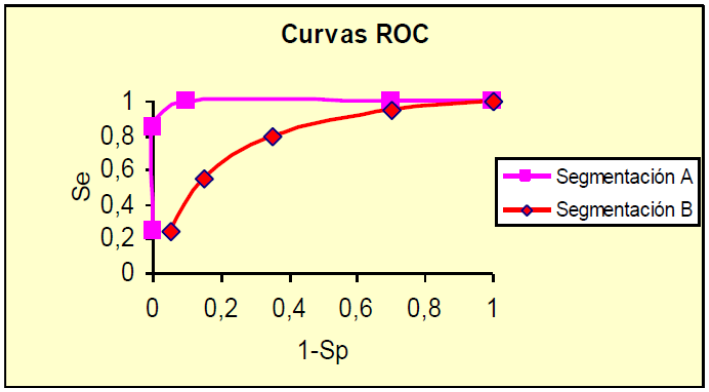

Figura 6. *Curvas ROC de dos modelos diferentes. Figura tomada de* [22].

Finalmente, es importante mencionar la importancia de la validación para trabajar con algoritmos supervisados. Siempre que se trabaja con este tipo de algoritmos, los datos deben dividirse en diferentes subconjuntos, de forma que aprendan de un subconjunto y se evalúen en otro. En caso contrario, se produciría lo que se conoce como sobreajuste, una situación de sobreentrenamiento del algoritmo en que se ajusta a características muy específicas y es incapaz de generalizar.

Entre las técnicas de validación destacan los métodos *hold-out* y *k-fold cross-validation*. *Hold-out* es cuando el conjunto de datos se divide en uno o más conjuntos de entrenamiento (*train set*) y en uno o más de evaluación (*test set*). El *train set* se utiliza para entrenar el modelo y el *test set* se

encarga de verificar cuán bien se desempeña cuando desconoce los datos de su entrada (el clasificador se evalúa con los datos del *test set* sin sus respectivas etiquetas). Así, cuando el clasificador devuelve sus predicciones, él mismo es capaz de comparar lo que ha predicho con la realidad (las etiquetas del *test set*). Estos resultados de evaluación de la validación son los que permiten ajustar los parámetros del clasificador y seleccionar el modelo que mejor se adapte a los datos [23].

Por su parte, el método *k-fold cross-validation* es un método de validación cruzada que consiste en dividir el conjunto de los datos en k subconjuntos, aplicando el método *hold- out* k veces. Para esta técnica se utiliza cada vez un subconjunto de datos distinto para validar el modelo, entrenado con los otros k-1 subconjuntos. Esta técnica tiene mayor utilidad cuando el conjunto de datos es pequeño, y tiene la ventaja de que todos los datos son utilizados tanto para entrenar como para validar [24]. En la Figura 7 se muestra un esquema ilustrativo de ambos métodos.

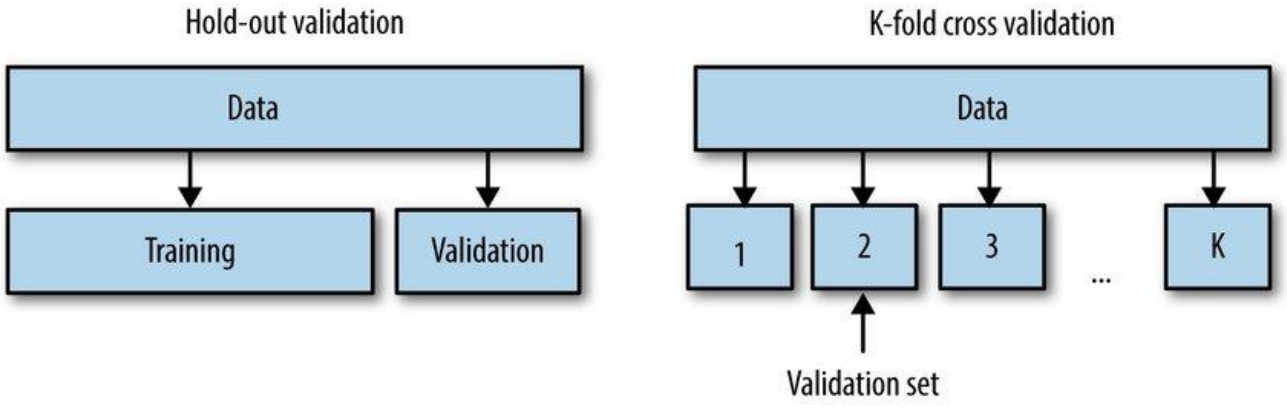

Figura 7. *Esquemas de funcionamiento de los métodos de validación hold-out y k-fold. Figura obtenida de* [23].

2.4. BIG DATA EN EL SECTOR SALUD

Del mismo modo que con ML, cuando se habla de IA se habla inevitablemente de macrodatos o *big data*, conjuntos de datos tan amplios y complejos que necesitan aplicaciones informáticas específicas para poder ser procesados. Es a través de este procesamiento que se obtienen características cualitativas y cuantitativas de los datos, de forma que a partir de ellos se pueden hacer análisis estadísticos avanzados y desarrollar modelos predictivos.

En el entorno de la salud se trabaja con grandes cantidades de datos heterogéneos que ofrecen al mismo tiempo información de los pacientes, de las enfermedades y de los propios centros sanitarios. Los datos clínicos se pueden obtener a partir de la historia clínica electrónica (HCE) del propio paciente, a partir de pruebas clínicas (pruebas de laboratorio, imágenes médicas, pruebas realizadas por un médico, etc.) o a través de determinados dispositivos portátiles de monitorización, como un pulsioxímetro o *wearables*. Estos datos contribuyen de forma significativa a optimizar la gestión de recursos en hospitales, así como el tratamiento y la atención a los pacientes.

A pesar de esto surge un problema, y es que, aunque los algoritmos de IA están automatizados, en cuanto respecta a la obtención y transmisión de los datos que se introducen a su entrada, muchas veces los datos no son válidos o no aportan suficiente valor. La generación de conocimiento es un proceso dinámico de síntesis, interpretación, integración y difusión de datos, por lo que el factor humano podría ser un inconveniente.

Es importante hacer mención también a los intereses subyacentes de empresas y a la dificultad de conseguir una completa colaboración de todas las partes implicadas en el proceso [25]. Además, una fracción significativa de los datos médicos disponibles continúa estando en papel (para mediados de 2017, el porcentaje de países de la Unión Europea con HCE no alcanzaba el 60% [26]) y todavía existen conflictos de privacidad [27].

En general, para una buena implementación de *big data* en los sistemas de salud se considera que existen desafíos en la agregación de los datos, en las políticas y procesos de los centros sanitarios y en la gestión de las empresas asociadas [28].

2.5. PROYECTO COVID DATA SAVE LIVES

El 25 de abril de 2020, HM Hospitales puso a libre disposición de la comunidad científica internacional una base de datos clínica y anonimizada con toda la información disponible sobre los pacientes tratados en sus centros hospitalarios por infección por el virus SARS-CoV-2. El objetivo del proyecto, llamado "COVID DATA SAVE LIVES", es el de facilitar datos clínicos a grupos de investigación, instituciones sanitarias, universidades y entidades científicas para avanzar en el conocimiento clínico, la adherencia a tratamientos y resultados asistenciales del manejo de pacientes durante la pandemia [29].

Esta base de datos destaca por incluir 2.307 HCE anonimizadas que recogen, además de datos demográficos, datos de las interacciones en el proceso de tratamiento de la COVID-19 tales como diagnósticos, tratamientos, ingresos, pasos por UCI, pruebas diagnósticos por imagen, resultados de laboratorio, altas o decesos, etc. [30]

A través de estos datos, HM Hospitales propuso de forma acorde a sus objetivos el uso de IA para obtener modelos predictivos de evolución y modelos epidemiológicos que permitiesen conocer más acerca de la evolución de la COVID-19.

2.6. ESTADO DEL ARTE

La predicción del progreso del estado de un paciente afectado por una enfermedad es un proceso complejo, más todavía si se trata de una enfermedad desconocida. Ante una patología nueva, la falta de conocimiento de la sintomatología completa y de indicadores clínicos sustanciales dificulta el trabajo de los profesionales sanitarios e impide conocer cómo va a evolucionar un paciente.

Surge ante este tipo de situaciones la necesidad de predictores que puedan ofrecer una estimación del estado del paciente introduciendo en su interfaz de entrada parámetros típicos de urgencia como la frecuencia cardíaca (FC) o la presión parcial de oxígeno. Se trata de desarrollar "calculadoras" médicas complejas que a través de macrodatos de pacientes faciliten la labor del personal sanitario, pues saber cómo se plantea su evolución es clave para elegir un tratamiento u otro.

Cuando se trata de predictores de COVID-19, la labor de estas calculadoras o sistemas consiste fundamentalmente en entrenar algoritmos con macrodatos de pacientes enfermos para detectar interrelaciones entre ciertos parámetros o biomarcadores y la defunción o no del paciente. Una vez entrenados dichos algoritmos, el sistema debe ser capaz de recibir datos clínicos de un paciente y predecir el modo en que se vaticina su evolución. En caso de que esta sea desfavorable, el médico o el personal sanitario pertinente solamente deberá limitarse a examinar cuál es el parámetro de riesgo que ha llevado a esta predicción y actuar en consecuencia.

La existencia de este tipo de sistemas no solo ayuda a la elección de tratamientos, sino que contribuye a amenizar los escenarios de congestión hospitalaria que se viven en situaciones como la descrita en los picos de pandemia. La toma de decisiones, el tratamiento y atención personalizada de los pacientes y otros factores como la correcta gestión de centros asistenciales se ven enormemente favorecidos con la aparición de estas herramientas [20].

2.6.1. Calculadora Médica COR+12

A mediados de abril de 2021, un grupo de profesionales liderado por el Servicio de Inmunología del Hospital Universitario 12 de Octubre desarrollaba una calculadora capaz de predecir el riesgo de mortalidad en pacientes ingresados con infección confirmada por SARS-CoV-2. Esta herramienta, la cual se ha denominado "COR+12", ha publicado los resultados de su investigación en el *Journal Allergy and Clinical Immunology* y se ha presentado en el marco del 2º Congreso Nacional Multidisciplinar COVID-19 de las Sociedades Científicas de España [31].

La calculadora se ha diseñado y validado con muestras de 1645 pacientes atendidos en el Hospital 12 de Octubre, el Hospital Ramón y Cajal y la Fundación Jiménez Díaz. En ella se valoran 5 variables para hacer la predicción: 3 parámetros de laboratorio (interleucina 6, enzima LDH y ratio neutrófilos/linfocitos), la edad y la saturación de oxígeno [31]. Estos parámetros son especialmente importantes a la hora de valorar la función pulmonar del paciente afectado y poder hacer una predicción.

Por otra parte, la COR+12 es totalmente gratuita y está ya disponible de forma *online* [32]. Como se observa en la Figura 8, la herramienta consiste en una sencilla interfaz donde solamente se deben introducir los parámetros de interés y ajustar los ejes de la gráfica que dará como resultado. Una vez hecha la predicción, COR+12 devuelve un resumen gráfico de la predicción de fallecimiento con un intervalo de confianza del 95% (véase la Figura 9). También se puede seleccionar un resumen numérico de los valores y un resumen del modelo, el cual consiste en un modelo de regresión logística.

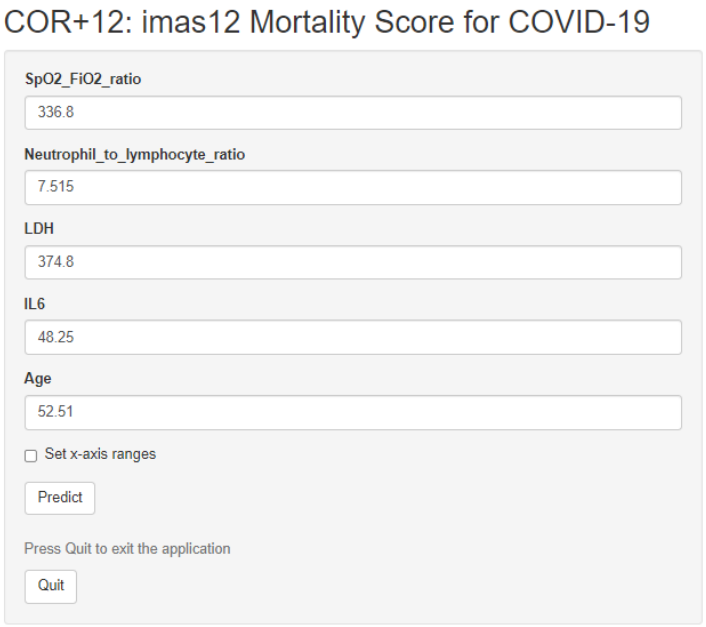

Figura 8. *Interfaz de la calculadora virtual COR+12. Figura obtenida desde la aplicación de la calculadora* [32].

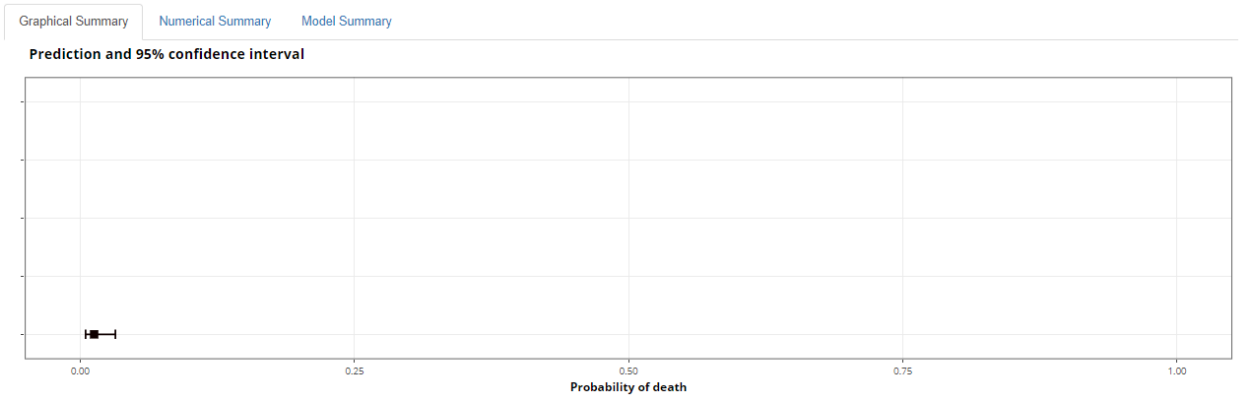

Figura 9*. Ejemplo de un resumen gráfico resultante de la predicción con COR+12. Figura obtenida desde la aplicación de la calculadora* [32].

Este modelo ha resultado ser muy útil para predecir qué pacientes se encuentran en alto riesgo de fallecimiento en su llegada a Urgencias o en los primeros días de hospitalización [31].

2.6.2. Sistema predictivo de Quirónsalud

Para mayo de 2021, los hospitales gestionados por el grupo hospitalario Quirónsalud incluyeron en sus sistemas un algoritmo de pronóstico de la evolución de personas ingresadas por COVID-19. Se trata de un algoritmo de ML que se actualiza periódicamente y que ya se ha incluido en las UCI y en las Unidades de Hospitalización de estos hospitales.

Los datos para su desarrollo provienen de los hospitales universitarios Fundación Jiménez Díaz, Rey Juan Carlos, Infanta Elena y General de Villalba, los cuales han aportado más de 15.000 casos de COVID-19 correspondientes a la primera ola de pandemia. Según explica el responsable de Big Data de estos centros sanitarios, "el sistema ofrece un patrón de comportamiento que permite prever la evolución, en términos de mortalidad y de empeoramiento (riesgo de ingreso en la UCI en las siguientes horas), de aquellos pacientes COVID hospitalizados que cumplen determinados criterios" [33].

Este algoritmo se basó en 20 variables entre las que se incluyeron información demográfica, antecedentes clínicos (diabetes, tabaquismo, hipertensión, enfermedades cardiovasculares…), toma de medicamentos, grupo sanguíneo, índice de masa corporal (IMC), ingreso previo en UCI o uso de ventilación mecánica. Sobre ellas se aplicaron árboles de decisión de hasta 4 niveles de complejidad y, a continuación, con un algoritmo tipo *Bayesian Ruleset* se pudo establecer un buen modelo de predicción. Este sistema sería capaz de predecir la probabilidad de que un paciente infectado por COVID-19 ingresara en UCI o falleciera [33].

Este sistema sobresale por su incorporación en tiempo real de los resultados a las bases de datos del hospital y la HCE de los pacientes, creando un aviso de predicción para ayudar a los profesionales sanitarios en la toma de decisiones [33]. Esto resulta también esencial para agilizar la actuación sobre los enfermos y mejorar la calidad asistencial.

2.6.3. Puntuación PANDEMYC

A mediados de agosto de 2020, un grupo de investigadores de la Universidad Rey Juan Carlos (URJC), la Universidad Complutense de Madrid (UCM), del Instituto de Salud Carlos III (ISCIII), del Instituto de Investigación Sanitaria Gregorio Marañón (IiSGM) y del Hospital Universitario Infanta Leonor crearon un modelo predictivo de mortalidad en pacientes hospitalizados por COVID-19 [34].

El sistema se basó en datos recogidos de 1968 pacientes ingresados en el Hospital Universitario Infanta Leonor durante la primera ola de pandemia. A partir de ellos se generó un modelo de puntuaciones basado en regresiones logísticas y árboles de clasificación que utiliza 9 parámetros de predicción al ingreso: edad, saturación de oxígeno, tabaquismo, creatinina sérica, linfocitos, hemoglobina, plaquetas, proteína C reactiva y sodio. Todos los parámetros tuvieron relevancia clínica similar a excepción de la edad, la saturación de oxígeno y la creatinina sérica, que resultaron ser de mayor peso y significación en el modelo de regresión [34].

PANDEMYC mostró en sus resultados una alta precisión diagnóstica, ofreciendo una herramienta útil de puntuación pronóstica confiable y fácil de utilizar, de gran utilidad en entornos de recursos limitados. Sin embargo, presenta la limitación de estar diseñada con datos de un único centro sin validación externa [34].

Del mismo modo que con las calculadoras virtuales, la puntuación PANDEMYC ofrece una interfaz *online* gratuita que pide a su entrada seleccionar entre distintos rangos de los parámetros de interés para hacer una estimación. Tras introducirlos y validar los datos, el sistema devuelve una puntuación y la probabilidad de fallecer durante el ingreso, asociada a dicha puntuación. En la Figura 10 se observa el modo en que se valora la puntuación para hacer una predicción.

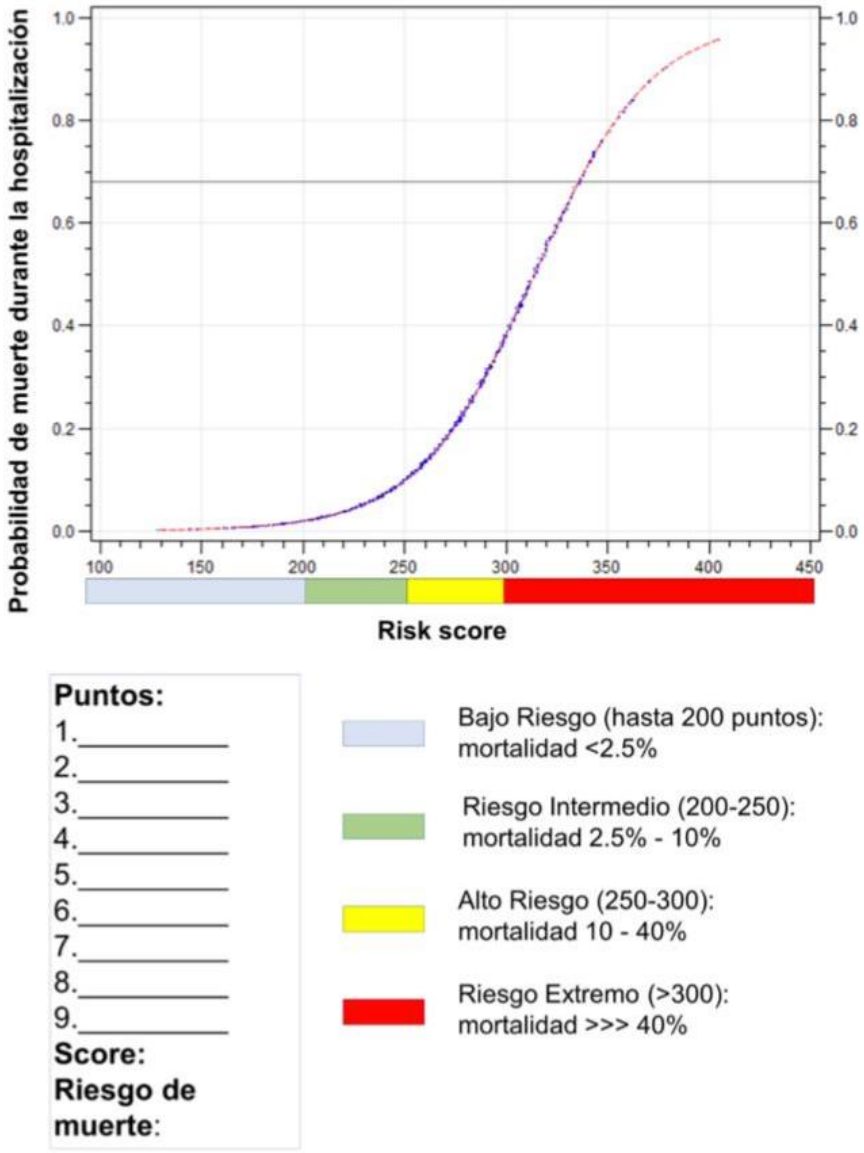

***Figura 10**. Fundamento en que se basa el sistema de puntuación PANDEMYC. Figura obtenida de* [34].

2.6.4. Puntuación COVAS

A nivel internacional, un grupo de médicos e investigadores del consorcio médico Kaiser Permanente de Estados Unidos (EE. UU.) desarrolló a finales de 2020 una herramienta para guiar las decisiones clínicas en Urgencias en pacientes con COVID-19. Esta herramienta, bautizada como "Puntuación COVAS", permite predecir con precisión la probabilidad la evolución de pacientes con síntomas de infección por SARS-CoV-2 [35].

Según la investigación, publicada en *The American Journal of Emergency Medicine,* se analizaron 26.600 casos de pacientes con sospecha de COVID-19 en su llegada a Urgencias. Se tomaron datos de 15 hospitales del grupo Kaiser Permanente comprendidos entre el 1 de marzo y el 30 de abril y se incluyeron como parámetros para predicción la edad, el sexo, el IMC y 7 comorbilidades (trastornos electrolíticos, arritmias cardíacas, otros trastornos neurológicos, pérdida de peso, insuficiencia cardíaca congestiva, coagulopatía y diabetes) [36].

Los resultados concluyeron con una buena precisión de la puntuación COVAS para predecir el fallecimiento o la descompensación respiratoria en un plazo de 7 días para pacientes que llegaran a Urgencias.

Es relevante mencionar que esta puntuación está diseñada para evaluar cómo puede evolucionar un paciente con sospecha de COVID-19, es decir, sin prueba diagnóstica confirmada. Esta herramienta responde a la necesidad de un sistema de ayuda para países y hospitales donde la sobresaturación y falta de recursos en oleadas de pandemia es importante y donde la toma de decisiones debe ser rápida.

EE. UU. es el lugar donde la pandemia ha tenido su mayor incidencia (mayor número de casos y de fallecidos a nivel mundial [8]), por lo que el desarrollo de la puntuación COVAS ha sido de significativa ayuda para evaluar el riesgo de los pacientes de Urgencias, agilizar la toma de decisiones y descongestionar los centros de atención sanitaria. Así, según afirman los propios creadores de la herramienta, se necesitan todavía investigaciones para validar más la puntuación, cambiando las poblaciones de pacientes y los entornos clínicos [36].

Según los creadores, la mejor aplicación hasta el momento de este sistema es la de identificar grupos de riesgo moderado que puedan evitar la hospitalización de manera segura, pero pudiendo beneficiarse de tratamientos domiciliarios como la monitorización de oxígeno y la medicación [36].

2.6.5. Colaboración entre HM Hospitales y el MIT Critical Data

Desde que HM Hospitales pusiera a disposición de la comunidad científica su *dataset* clínico, más de un centenar de instituciones de hasta 32 países diferentes han mostrado su interés por él. Existen distintos grupos de investigación y corporaciones sanitarias trabajando con estos datos y que ya han presentado proyectos a la espera de ser aceptados [29].

Destaca en este proyecto el acuerdo de colaboración entre HM Hospitales y el *Massachusets Institute of Technology (*MIT) para encontrar algoritmos de predicción derivados del COVID DATA SAVE LIVES. Con la asociación se pretende resolver los misterios que entraña la COVID-19, así como mejorar el tratamiento de pacientes, perfeccionar la predicción de la evolución de cada paciente particular y contribuir a una mejor gestión de su impacto en la población [37].

Hasta el momento se conoce que dicha colaboración ha centrado sus esfuerzos en desarrollar algoritmos de aprendizaje profundo o *deep learning* para mejorar la predicción de resultados de pacientes con COVID-19 utilizando datos clínicos de estos e imágenes de placas de tórax. También se conoce que se está realizando un estudio retrospectivo con este *dataset* que investiga el impacto de fármacos inhibidores de la interleucina 6 en pacientes ingresado [37].

3. MATERIALES

Se exponen en este capítulo los distintos materiales empleados para poder alcanzar los objetivos propuestos.

Índice de contenidos

3.1. BASE DE DATOS DEL PROYECTO COVID DATA SAVE LIVES

En mayo de 2020, HM Hospitales cedió al grupo de investigación CVBLab del Instituto de Investigación en Bioingeniería (I3B) de la Universidad Politécnica de Valencia su *dataset* para trabajar en un modelo automático de predicción.

La información en dicho *dataset* se encuentra organizada en tablas según su contenido, donde cada paciente individual posee un identificador (ID) único que recoge sus datos. Este identificador no tiene relación alguna con el identificador real del ingreso al hospital de los pacientes, pues los datos son completamente anónimos.

La primera tabla (Tabla 1ª a partir de ahora), es la tabla principal y recoge datos propios del ingreso y del paciente (edad y sexo), datos de la urgencia previa si ha habido, su estancia en UCI si ha habido, los registros del primer y último conjunto de constantes de la urgencia, el diagnóstico de urgencia, la unidad desde donde se le deriva y el motivo de alta del paciente (alta voluntaria, fallecimiento, domicilio, traslado al hospital o traslado a un centro sociosanitario). Incluye datos de 2307 pacientes y recoge entre las constantes la temperatura, la frecuencia cardíaca (FC), la glucemia, la saturación de oxígeno, la tensión arterial mínima (PA mínima) y la tensión arterial máxima (PA máxima).

La segunda tabla (Tabla 2ª a partir de ahora), recoge información de la medicación pautada y administrada durante el ingreso (no durante su estancia en UCI). Incluye el nombre comercial del fármaco, la dosis media diaria administrada, la duración del tratamiento y la descripción de su categoría ATC5 y ATC7 de la clasificación ATC (*Anatomical, Therapeutical, Chemical classification system*). Se encuentran en esta tabla 60.460 registros.

En la tercera tabla (Tabla 3ª a partir de ahora), se recoge información sobre los registros de las constantes durante todo el ingreso de los pacientes, no recogiéndose así las de UCI. Se incluye en esta

tabla la fecha y hora del registro, siendo las constantes las mismas que en la Tabla 1. Se recogen un total de 55.515 registros en esta tabla.

En la cuarta tabla (Tabla 4ª a partir de ahora), la tabla de laboratorio, se recogen los registros de los resultados de peticiones de laboratorio realizadas durante el ingreso y en la urgencia previa, si es que hubo. Se incluyen también en esta tabla los resultados de cada prueba. En total, la Tabla 4ª cuenta con 396.055 registros.

En la quinta de las tablas proporcionadas (Tabla 5ª a partir de ahora), se recogen los registros de la codificación de la urgencia según la clasificación internacional CIE10 en la última versión distribuida en aquel momento (no incluía COVID-19). Se recogen hasta 12 diagnósticos (uno principal y 11 posibles diagnósticos secundarios) y hasta 5 procedimientos. Esta tabla recoge 1988 registros.

Finalmente, en una sexta tabla (Tabla 6ª a partir de ahora) se recogen los registros de la codificación del ingreso del mismo modo que en la Tabla 5ª para la urgencia, incluyendo también neoplasias e indicadores POA (*Present On Admission*), relacionados con la clasificación CIE10. Esta tabla recoge 1775 registros.

Adicionalmente se facilitó al CVBLab el conjunto de imágenes de tórax tomadas a los distintos pacientes y asociadas a ellos a través de su ID. El total resultó en 5481 imágenes.

En el Anexo I se han incluido tablas resumen que recogen el contenido de cada una de las tablas y resumen sus parámetros según fueron proporcionados por HM Hospitales.

3.2. SOFTWARE

Los datos de las distintas tablas se proporcionaron a CVBLab como archivos CSV (*Comma Separated Values*), un tipo de archivo de texto en que la información se encuentra separada por comas formando una especie de tabla en filas y columnas. Para la visualización de dichos datos se ha empleado el software de hojas de cálculo Microsoft Excel.

Para el tratamiento de los datos, su análisis y el posterior desarrollo de los distintos algoritmos se ha utilizado el programa MATLAB® v.R2020b, de MathWorks. MATLAB es una plataforma de programación y cálculo numérico que combina un entorno de escritorio perfeccionado para el análisis iterativo y los procesos de diseño con un lenguaje propio que expresa las matemáticas de matrices y *arrays* directamente (lenguaje M). MATLAB permite desde explorar, modelar y visualizar datos hasta desarrollar algoritmos y entrenarlos [38].

Para el desarrollo de los modelos se ha utilizado la aplicación *Classification Learner* que ofrece MATLAB. Esta herramienta, accesible a través del comando *classificationLearner*, permite entrenar distintos modelos para clasificar datos, de modo que, tras su entrenamiento, el modelo desarrollado es capaz de hacer predicciones para nuevos datos.

4. MÉTODOS

Utilizando los materiales descritos, en el presente capítulo se recogen los distintos procedimientos llevados a cabo para desarrollar un algoritmo de ML capaz de clasificar un conjunto de pacientes con COVID-19 en función de su evolución. Para ello, se divide esta sección en dos partes diferenciadas, siendo cada una de ellas vital para asegurar un buen resultado.

Índice de contenidos

4.1. ANÁLISIS EXPLORATORIO DE LOS DATOS

El primer paso en el proceso para elaborar un algoritmo de ML es la recolección de datos. Dado que se trabaja con una base de datos ya configurada y estructurada, el primer paso de este proyecto consiste en analizar la calidad de los datos con que se va a trabajar.

4.1.1. Limpieza de los Datos

Se conocen como *dirty data* aquellos datos que son erróneos, inconsistentes o incompletos, especialmente cuando se encuentran en una base de datos. Estos datos significan un inconveniente a la hora de desarrollar cualquier proyecto y llevan a la pérdida de información.

La aparición de *dirty data* en una base de datos se asocia a distintos factores, entre los que se encuentran el volumen de los datos, la introducción de errores en el registro, los silos de información, la falta de información y los datos falseados [21]. En el entorno de salud, todos estos factores tienen influencia en la información contenida en bases de datos clínicas, ya que la introducción de los datos es casi siempre manual y múltiples sistemas accediendo a un mismo programa pueden generar incoherencias en los datos.

Para tratar estos datos el primer paso consiste en su detección. Se trata de detectar los campos que se encuentran vacíos o que contienen datos inconsistentes, *outliers* que se alejan de un valor dentro de un intervalo normal. A continuación, la opción correcta consiste en corregir los datos,

asignándoles valores probables o la media, si se trata de un valor numérico. Sin embargo, esto da lugar a errores posteriores, por lo que la mayoría de los especialistas decide eliminar el registro completo cuando aparecen este tipo de datos.

4.1.2. Selección de los Datos

Una vez los datos estén limpios, fijando como objetivo la obtención de un algoritmo que prediga la evolución de un paciente a su entrada a Urgencias con una prueba de COVID-19 positiva, se seleccionan aquellos datos que puedan ser relevantes para el desarrollo del modelo (sus predictores) y se descartan aquellos que no tengan una relación significativa.

4.1.3. Análisis Estadístico de los Datos

Ya reducido el *dataset* de forma que se tenga en un mismo archivo toda la información relevante, se procede a realizar un análisis estadístico de los datos que proporcione información acerca de las variables y de la correlación entre ellas.

Para ello, se analizan los datos de las distintas tablas a fin de conocer lo máximo posible acerca de la información contenida en ellas. Se divide en datos numéricos y datos categóricos. Ambos se examinan de igual modo, pero adicionalmente se obtienen de los datos numéricos medidas características como la media, la desviación típica, la mediana, la moda y el rango.

- **Media:** devuelve el valor medio de un conjunto de datos.

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{5}$$

- **Desviación típica**: cuantifica la variación o la dispersión de un conjunto de datos numéricos. Es útil porque indica cuánto pueden alejarse de la media los valores de un conjunto de datos.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}} \tag{6}$$

- **Mediana**: es el valor central en un conjunto ordenado de datos. En estadística da información acerca de cómo se distribuyen los datos.

- **Moda**: es el valor que más se repite dentro de un conjunto de datos.
- **Rango**: muestra el intervalo de valores entre los cuales se encuentra el conjunto de los datos.

Tras dicho análisis, el siguiente paso consiste en obtener correlaciones entre variables. Para ello se normalizan los datos y se obtiene a través de Matlab un mapa de correlaciones a través del comando *gplotmatrix*.

4.2. CONSTRUCCIÓN DE UN MODELO DE APRENDIZAJE AUTOMÁTICO

Una vez el conjunto de datos ha quedado reducido y se conocen las interrelaciones entre variables, el procedimiento se reduce a estudiar los datos correspondientes a las variables de interés utilizando la aplicación disponible en MATLAB *Classification Learner*.

Esta herramienta permite entrenar múltiples tipos de clasificadores, agrupados según se incluyan dentro de árboles de decisión, análisis discriminante lineal, regresión logística, clasificación bayesasiana ingenua, máquinas de vectores de soporte, clasificación de K vecinos más próximos o métodos conjuntos.

- **Árboles de decisión *(Decision Tree Algorithms)***

Los árboles de decisión son un tipo de algoritmos de aprendizaje supervisado que a través de reglas sencillas inferidas a partir del conjunto de datos de entrenamiento consiguen establecer una clasificación. Sirven para datos numéricos y categóricos.

En MATLAB se encuentran disponibles 3 algoritmos de clasificación de árboles de decisión, *coarse tree, medium tree* y *fine tree.* Los 3 modelos tienen una velocidad de predicción rápida, un uso de memoria bajo y una fácil interpretabilidad. La diferencia entre ellos radica en la flexibilidad del modelo (el número de divisiones que admite): *coarse tree* admite hasta 4 divisiones, *medium tree* hasta 20 y *fine tree* alcanza las 100 [39]. Su principal inconveniente es que pueden caer en el sobreajuste y en muchas ocasiones su precisión no es suficiente.

- **Análisis discriminante**

Este tipo de análisis se centra en desarrollar funciones discriminantes a partir de la combinación lineal de variables independientes para posteriormente discriminar entre las categorías de la variable dependiente. Este método se utiliza para analizar datos cuando la variable dependiente es categórica y las variables independientes tienen naturaleza de intervalo [40].

Dentro de MATLAB existen dos tipos de clasificadores, lineal (*linear discriminant)* y cuadrático *(quadratic discriminant*). Ambos obtienen predicciones de forma rápida, son fáciles de interpretar y tienen una baja flexibilidad del modelo. La diferencia entre ambos radica en su uso de memoria, siendo mayor para los clasificadores cuadráticos [39].

- **Regresión logística**

La regresión logística es un tipo de análisis de regresión que se utiliza para predecir el resultado de una variable categórica en función de variables independientes. Modela la probabilidad de que una muestra pertenezca a una clase en función de otras variables.

En MATLAB la regresión logística es un algoritmo de predicción rápido, con uso medio de memoria y una flexibilidad baja. Es especialmente popular cuando se busca clasificar datos en dos clases, pues es bastante fácil de interpretar [39].

- **Clasificación bayesiana ingenua *(Naive Bayes Classifiers)***

Este tipo de clasificadores se fundamenta en el teorema de Bayes (probabilidad condicional) y tiene la particularidad de que asume independencia entre las variables sobre las que hace la predicción.

En la herramienta de clasificación de MATLAB este método tiene dos subvariantes, g*aussian naive bayes* y *kernel naive bayes*. Mientras que la clasificación gaussiana *(gaussian naive bayes)* tiene una velocidad media, un uso de memoria entre bajo y medio y una flexibilidad baja, la clasificación bayesiana utilizando estimaciones de densidad del núcleo *(kernel naive bayes)* tiene una velocidad de predicción baja, un uso de memoria medio y una flexibilidad media. Ambos tienen fácil interpretabilidad [39].

- **Máquinas de vectores de soportes (*Support Vector Machines*)**

Las máquinas de vectores de soportes (por sus siglas en inglés, SVM) son conjuntos de algoritmos de aprendizaje supervisado que dividen el conjunto de datos en distintas clases según hiperplanos de separación. La clasificación que realizan depende del área del espacio formada por dichos hiperplanos en donde caiga cada muestra. Si la clasificación se hace para 2 clases, se dice que el clasificador es binario; en caso de ser más clases, se habla de un clasificador multiclase.

En MATLAB existen diferentes tipos de SVM que se diferencian según se especifica en la Tabla 2. Cada tipo de clasificador utiliza una función del núcleo diferente para definir el hiperplano de separación.

***Tabla 2**. Comparación de los distintos tipos de SVM disponibles en el Classification Learner de MATLAB. Tabla obtenida de* [39].

Tipo de clasificador	Velocidad de predicción	Uso de memoria	Interpretabilidad	Flexibilidad del modelo
SVM Lineal	Binario: Rápida Multiclase: Media	Medio	Fácil	Baja
SVM Cuadrático	Binario: Rápida Multiclase: Lenta	Binario: Medio Multiclase: Elevado	Difícil	Media
SVM Cúbico	Binario: Rápida Multiclase: Lenta	Binario: Medio Multiclase: Elevado	Difícil	Media

SVM *Fine Gaussian*	Binario: Rápida Multiclase: Media	Binario: Medio Multiclase: Elevado	Difícil	Elevada
SVM *Medium Gaussian*	Binario: Rápida Multiclase: Lenta	Binario: Medio Multiclase: Elevado	Difícil	Media
SVM *Coarse Gaussian*	Binario: Rápida Multiclase: Lenta	Binario: Medio Multiclase: Elevado	Difícil	Baja

- **Clasificación por vecinos más cercanos *(K-Nearest Neighbor Classifiers)***

La clasificación por vecinos más cercanos (por sus siglas en inglés, KNN) clasifica una muestra en la misma clase que sus K vecinos más cercanos, siendo K un número entero predefinido.

De nuevo aparecen en MATLAB distintos subtipos de clasificadores KNN, diferenciados cada uno por la métrica empleada a la hora de establecer la distancia a los vecinos más próximos. La aplicación incluye *fine KNN, medium KNN, coarse KNN, cosine KNN, cubic KNN* y *weighted KNN*. Del mismo modo que con SVM, las diferencias entre cada subtipo se encuentran recogidas en la Tabla 3. Es importante mencionar que la flexibilidad en este clasificador hace referencia al número de vecinos.

***Tabla 3**. Comparación de los distintos tipos de KNN disponibles en el Classification Learner de MATLAB. Tabla obtenida de* [39].

Tipo de clasificador	Velocidad de predicción	Uso de memoria	Interpretabilidad	Flexibilidad del modelo
Fine KNN	Media	Media	Difícil	Elevada (1 vecino)
Medium KNN	Media	Media	Difícil	Media (10 vecinos)
Coarse KNN	Media	Media	Difícil	Baja (100 vecinos)
Cosine KNN	Media	Media	Difícil	Media (10 vecinos)
Cubic KNN	Lenta	Media	Difícil	Media (10 vecinos)
Weighted KNN	Media	Media	Difícil	Media (10 vecinos)

- **Métodos conjuntos *(Ensemble Classifiers)***

Los métodos conjuntos mezclan resultados de varios clasificadores con más bajo poder de predicción para conseguir un modelo conjunto de alta calidad. Al utilizar múltiples algoritmos de aprendizaje consiguen reducir errores y obtener un mejor rendimiento predictivo.

Dentro de estos métodos se encuentran en el *Classification Learner* los métodos *boosted trees, bagged trees, subspace discriminant, subspace KNN* y *RUSBoosted trees*. Sus diferencias quedan recogidas en la Tabla 4.

Tabla 4. *Comparación de los distintos tipos de métodos conjuntos disponibles en el Classification Learner de MATLAB.Tabla obtenida de* [39].

Tipo de clasificador	Velocidad de predicción	Uso de memoria	Interpretabilidad	Flexibilidad del modelo
Boosted trees	Rápida	Bajo	Difícil	Media - alta
Bagged trees	Media	Alto	Difícil	Alta
Subspace discriminant	Media	Bajo	Difícil	Media
Subspace KNN	Media	Medio	Difícil	Media
RUSBoosted trees	Rápida	Bajo	Difícil	Media

Teniendo en cuenta que el *dataset* incluye tanto datos numéricos como categóricos, se prueban distintos clasificadores en busca del que mejor resultado proporcione. Para ello se realiza el entrenamiento y validación con cada uno de los algoritmos y se selecciona el que mayor precisión proporcione. Todos los algoritmos se prueban tanto con validación *hold-out* como *k-fold*.

En caso de que la predicción no ofrezca buenos resultados y suponiendo que la base de datos sobre la cual se realiza el análisis cuenta con datos de calidad, debe repetirse el proceso en busca de mejores predictores que discriminen mejor entre clases.

Finalmente, se valida el modelo con datos externos que el algoritmo desconozca, a fin de asegurar el éxito en el desarrollo del mismo. Sin embargo, este último paso no se puede llevar a cabo en el presente proyecto por falta de datos de centros distintos a HM Hospitales, de modo que se deja esta validación para ser probada en proyectos futuros.

5. RESULTADOS

Siguiendo el mismo esquema que se plantea en el capítulo anterior, se recogen en este capítulo los resultados obtenidos de aplicar las técnicas propuestas en este.

Índice de contenidos

5.1. ANÁLISIS EXPLORATORIO DE LOS DATOS

Para poder llevar a cabo un correcto análisis de los datos proporcionados por HM Hospitales, se realizó una selección previa de los datos de interés, tomando como criterio la relevancia que estos pudiesen tener para determinar la defunción o no de un paciente que es ingresado por COVID-19. Considerando los objetivos planteados, se descartaron las imágenes de tórax para el desarrollo del clasificador.

Por otra parte, si se considera la necesidad de desarrollar un clasificador capaz de predecir la evolución de los pacientes con COVID-19 cuando llegan a la urgencia, se podrían descartar por el momento los datos recogidos en la Tabla 3ª, pues estos incluyen los datos durante el ingreso, estando los datos de la urgencia recogidos en la Tabla 1ª. Sucede del mismo modo con la Tabla 6ª y la Tabla 5ª.

La Tabla 1ª es la que más datos contiene en cuanto a características de los pacientes y datos en Urgencias. Si bien no recoge datos de patologías previas que podrían ser significativas para definir un perfil de riesgo, se recogen en ella las constantes vitales de los pacientes y la variable objetivo del clasificador a desarrollar: el motivo de alta.

5.1.1. Limpieza de los Datos

5.1.1.1. Tabla 1ª

En la Tabla 1ª se encuentra la variable objetivo en torno a la cual giraría el funcionamiento del futuro predictor. Es crucial conocer el motivo del alta de los pacientes para poder predecir su evolución, de modo que todos aquellos registros que no contenían información de esta variable se eliminaron automáticamente. Se eliminaron 259 registros.

Se analizó a continuación la cantidad de datos vacíos de cada variable, tomándose como criterio de exclusión para el estudio que las variables tuviesen más de un 60% de datos vacíos. Se eliminaron del análisis la glucemia tomada a la entrada a Urgencias y los datos relativos al ingreso en UCI (véase la Tabla 5).

Tabla 5. *Datos vacíos según porcentaje de las distintas variables recogidas en la Tabla 1ª. En esta tabla se excluye "Motivo de alta" por haberse eliminado anteriormente todos los registros vacíos.*

Datos vacíos	
Edad	0%
Sexo	0%
Diagnóstico al ingreso	0%
Fecha de ingreso	0%
Fecha de alta	0%
Fecha de ingreso en UCI	92,77%
Fecha de alta en UCI	92,77%
Días en UCI	92,77%
Fecha de admisión en Urgencias	3,13%

Hora de admisión en Urgencias	3,13%
Especialidad de donde deriva el ingreso	3,13%
Diagnóstico de Urgencias	3,13%
Destino	3,13%
Hora primera toma de constantes de Urgencias	18,51%
Hora última toma de constantes de Urgencias	3,13%
Primera toma de temperatura	23,75%
Última toma de temperatura	20,98%
Primera toma de FC	22,93%
Última toma de FC	17,95%
Primera toma de glucemia	98,96%
Última toma de glucemia	23,75%
Primera toma de saturación de oxígeno	21,63%
Última toma de saturación de oxígeno	16,52%
Primera toma de PA máxima	36,50%
Última toma de PA máxima	32,51%
Primera toma de PA mínima	36,54%
Última toma de PA mínima	32,55%

Durante este análisis se observaron inconsistencias en las variables “Fecha de admisión en Urgencias” y “Última toma de glucemia”. Se observó una coincidencia completa de estas variables con “Fecha de ingreso” y “Primera toma de temperatura”, respectivamente. En el primer caso, la inconsistencia se debía, como especificaba el propio grupo HM Hospitales, a la condición de ingreso hospitalario que recibieron los pacientes que llegaban a Urgencias con sospecha de COVID-19, realizándose ambos procesos en la misma fecha. En el segundo caso, se observó que los datos de temperatura a la entrada a Urgencias habían sido copiados erróneamente en la última toma de glucemia. Ya fuera por falta de información adicional o por error, ambas variables se excluyeron.

Se observó también en este punto que, obviando los valores vacíos de las variables, todos los datos de constantes vitales tomados a la entrada a Urgencias tenían valores idénticos a la salida. Es decir, en todos los registros con valores distintos a cero en la primera toma y la última, los valores numéricos eran exactamente iguales.

Excluir datos ante esta situación era una decisión controversial, pues existían registros donde se había registrado el valor de la primera toma y no de la última o viceversa. Esto significa que no se

podía conocer a priori qué datos fueron copiados de cuáles. Sin embargo, a fin de satisfacer el objetivo de generar un clasificador capaz de realizar una predicción a la entrada a Urgencias, se consideró que los datos correctos se tomaron a la entrada y se excluyeron para la elaboración del clasificador las variables últimas de la urgencia.

Finalmente, se estudió que los datos de las constantes vitales se encontrasen dentro de rangos normales o habituales en la práctica clínica, según se establece a continuación.

- **Valores normales de temperatura**

La temperatura corporal es una constante que varía ampliamente dependiendo de la zona del cuerpo en que se tome (oral, rectal, axilar, oído, escáner infrarrojo), la edad o el sexo. Considerando valores normales para adultos y teniendo en cuenta que en el momento en que se realizó la toma de temperatura a los pacientes no existían tantas precauciones higiénico-sanitarias en Urgencias, se consideran como referencia los valores en axila. Se consideraron entonces valores normales los comprendidos entre 35,2ºC y 36,9ºC [41].

Por otra parte, se conoce que valores comprendidos entre 30ºC y 35ºC indican hipotermia, mientras por encima de 41ºC indican casos graves de fiebre en que debe revertirse cuanto antes el aumento de la temperatura [41]. Si la temperatura alcanza los 42ºC el paciente entra en coma y a los 43ºC muere por efecto de la desnaturalización de las proteínas [42].

Se tomaron de acuerdo a esto como umbrales inferior y superior las temperaturas 30ºC y 42ºC. No se detectó ninguna inconsistencia en los datos.

- **Valores normales de FC**

Se entiende que valores de FC en reposo por debajo de 60 latidos por minuto (lpm) corresponden a bradicardia y por encima de 100 lpm a taquicardia [43]. Sin embargo, existen excepciones a esta regla que dependen de múltiples factores externos.

Teniendo en cuenta lo mencionado, se estudiaron los casos mínimos y máximos de estos valores para poder detectar anomalías en los datos. No se descartaron datos de ningún paciente.

- **Valores normales de saturación de oxígeno (presión arterial parcial de oxígeno, PaO_2)**

Se considera que un paciente tiene una PaO_2 disminuida cuando esta constante baja de 80 mmHg. Cuando este valor cae por debajo de 60 mmHg, se considera insuficiencia respiratoria, significando a niveles inferiores la necesidad inmediata de oxigenoterapia [44]. Por otra parte, valores de PaO_2 por encima de 100 mmHg indican hiperoxemia (nivel de oxígeno en sangre elevado), pudiendo ser este trastorno leve (100 – 200 mmHg), moderado (200 – 300 mmHg) o elevado (< 300 mmHg) [45].

Con esta información y los datos contenidos en la base de datos se marcaron los umbrales para determinar posibles *outliers*, considerando la posible variabilidad asociada a esta medida. Solamente se pudo excluir a un paciente cuya PaO_2 era completamente anormal, de 10 mmHg.

- **Valores normales de tensión arterial**

En cuanto respecta a valores normales de la tensión o presión arterial (PA), deben considerarse la PA mínima (diastólica) y la PA máxima (sistólica). La categoría de la PA puede definirse siguiendo los criterios establecidos por el *American College of Cardiology/American Heart Association Hypertension* (2017) en la Tabla 6 [46].

Se tomaron umbrales relativos a ambas presiones y se encontraron ciertas anomalías en los datos. Se detectaron y eliminaron 2 registros con valores totalmente fuera de un rango posible. Adicionalmente, se eliminaron aquellos registros en que los valores de presión diastólica eran superiores a los de la presión sistólica (6 pacientes).

Tabla 6. *Categorías de PA según los valores de presión sistólica y de presión diastólica. Tabla obtenida de* [46].

Categoría de la presión arterial	Presión sistólica (mmHg)		Presión diastólica (mmHg)
Normal	< 120	y	< 80
Elevada	120 – 129	y	< 80
Hipertensión Nivel 1	130 – 139	o	80 – 89
Hipertensión Nivel 2	≥ 140	o	≥ 90
Hipertensión Nivel 3	> 180	y / o	≥ 120

Tras el primer proceso de limpieza de datos de la Tabla 1ª, el número total de pacientes se redujo a 2039.

5.1.1.2. Tabla 2ª

En esta tabla se observan los distintos tratamientos realizados a 2300 de los 2307 pacientes a los que se tomaron datos. No se requirió ningún tipo de limpieza en esta tabla.

5.1.1.3. Tabla 4ª

Se observa en esta tabla un conjunto de datos caótico en que aparecen anotaciones mezcladas con otras variables. Se encuentran hasta 400 pruebas diferentes de laboratorio entre los más de 396.000 registros, habiéndosele realizado a algunos pacientes incluso 4400 pruebas.

Si bien no resulto necesaria una limpieza de esta tabla, la heterogeneidad de los datos que contiene sugería reconsiderar su inclusión en un clasificador del nivel del que se pretendía realizar.

5.1.1.4. Tabla 5ª

En la última tabla a analizar se observó que solamente se habían tomado datos del diagnóstico y procedimientos a algunos de los pacientes. Entre los datos correspondientes a esta tabla se observaron datos vacíos en altos porcentajes (véase Tabla 7), lo que sugería considerar solamente el valor correspondiente a "Diagnóstico Principal" a la hora de realizar el clasificador.

***Tabla 7**. Datos vacíos según porcentaje de las distintas variables recogidas en la Tabla 5ª.*

Datos vacíos	
Diagnóstico principal	0%
Diagnóstico número 2	59,59%
Diagnóstico número 3	85,81%
Diagnóstico número 4	94,01%
Diagnóstico número 5	97,53%
Diagnóstico número 6	98,89%
Diagnóstico número 7	99,30%
Diagnóstico número 8	99,65%
Diagnóstico número 9	99,85%
Diagnóstico número 10	99,95%
Diagnóstico número 11	100%
Procedimiento número 1	92,50%
Procedimiento número 2	97,38%
Procedimiento número 3	99,45%
Procedimiento número 4	100%
Procedimiento numero 5	100%

5.1.2. Selección de los Datos

Se analizó en este apartado el valor de los datos limpios a la hora de formar parte de las variables predictoras que configurarían posteriormente el predictor. Se realizó de este modo una selección de las variables a fin de que se mantuviesen solamente aquellas que pudieran servir para conocer la evolución de un paciente con COVID-19.

5.1.2.1. Tabla 1ª

Se excluyeron del estudio las variables relativas a fechas y horas. Se excluyeron además las variables "Diagnóstico al ingreso" y "Destino", ya que en todos los casos se trataba de un ingreso originado por COVID-19. Por no tener mayor relevancia, se descartó también la variable "Especialidad de donde deriva el ingreso".

5.1.2.2. Tabla 2ª

Del mismo modo que en la Tabla 1ª, se excluyeron las variables relativas a fechas. Se eliminaron de este modo las variables "Inicio del tratamiento" y "Fin del tratamiento". Por otra parte, resultaba especialmente interesante considerar el tratamiento con la clasificación ATC7, pues es la clasificación más completa de entre las ofrecidas en el *dataset*, de modo que solamente se reservó esta variable asociada al ID del paciente en el desarrollo del clasificador.

5.1.2.3. Tabla 4ª

Por la heterogeneidad mencionada anteriormente, las pruebas de laboratorio correspondientes a la Tabla 4ª quedaron descartadas del estudio.

5.1.2.4. Tabla 5ª

Teniendo en cuenta que los registros de diagnósticos realizados sobre los pacientes no incluían en el momento de su ingreso una codificación en CIE10 para COVID-19, los datos recogidos en esta tabla no tienen mayor valor clínico que una aproximación médica hacia una sintomatología parecida a la de la COVID-19. Dado que además los valores para los procedimientos y para los distintos diagnósticos secundarios que se proporcionan en esta tabla son muy bajos, la Tabla 5ª se descartó para su uso en el desarrollo del clasificador.

5.1.3. Preprocesado Final de los Datos

Previamente a analizar y utilizar los datos, cabía tratar todos los campos correspondientes a variables que se encontraran vacíos. Como se mencionó anteriormente, cabe la posibilidad de realizar una corrección de estos datos a través de distintos métodos como la sustitución del valor vacío por la media aritmética; sin embargo, se optó por eliminar los registros de todos aquellos pacientes con algún campo vacío. Se evitó así introducir errores adicionales a la vez que se trabajaba con datos completos y reales.

En la Tabla 1ª, el número de pacientes quedó reducido a 1192. La Tabla 2ª, por coherencia con la Tabla 1ª en cuanto al uso de medicamentos de cada paciente, quedó también reducida a 1192 pacientes (28.957 registros). Se eliminaron de esta segunda tabla los registros "sin clasificar", quedando un total de 28.394 registros.

En la Figura 11, situada en la siguiente página, se recoge un diagrama de flujo representativo del proceso de limpieza, selección y preprocesado de los datos originales.

5.1.4. Análisis Estadístico de los Datos

Sobre las dos tablas resultantes del cribado anterior se realizó en este punto un análisis estadístico que proporcionase información acerca de los datos seleccionados. Esta información permite conocer correlaciones entre variables y, a posteriori, la posibilidad de afianzar conclusiones acerca de los resultados del clasificador.

5.1.4.1. Tabla 1ª

Los datos correspondientes a los resultados del análisis de datos se dividieron en datos numéricos y datos categóricos. A continuación, se detalla el análisis llevado a cabo sobre los datos categóricos y en la Tabla 8 se recogen los resultados correspondientes al análisis estadístico que se realizó sobre los datos numéricos.

- **Sexo**

En la Tabla 1ª se recogen datos de 720 hombres y 472 mujeres. El porcentaje total de mujeres en el estudio solamente representa el 39,6%.

- **Motivo de alta**

En los datos de estudio solamente 197 de los 1194 pacientes fallecieron (un 16,5%). Los demás pacientes, según orden de cantidad, salieron del hospital con motivo de alta domiciliaria (78,14%), traslado al hospital fuera de zona COVID (2,85%), traslado a otro centro sociosanitario (2,35%) y alta voluntaria (0,17%).

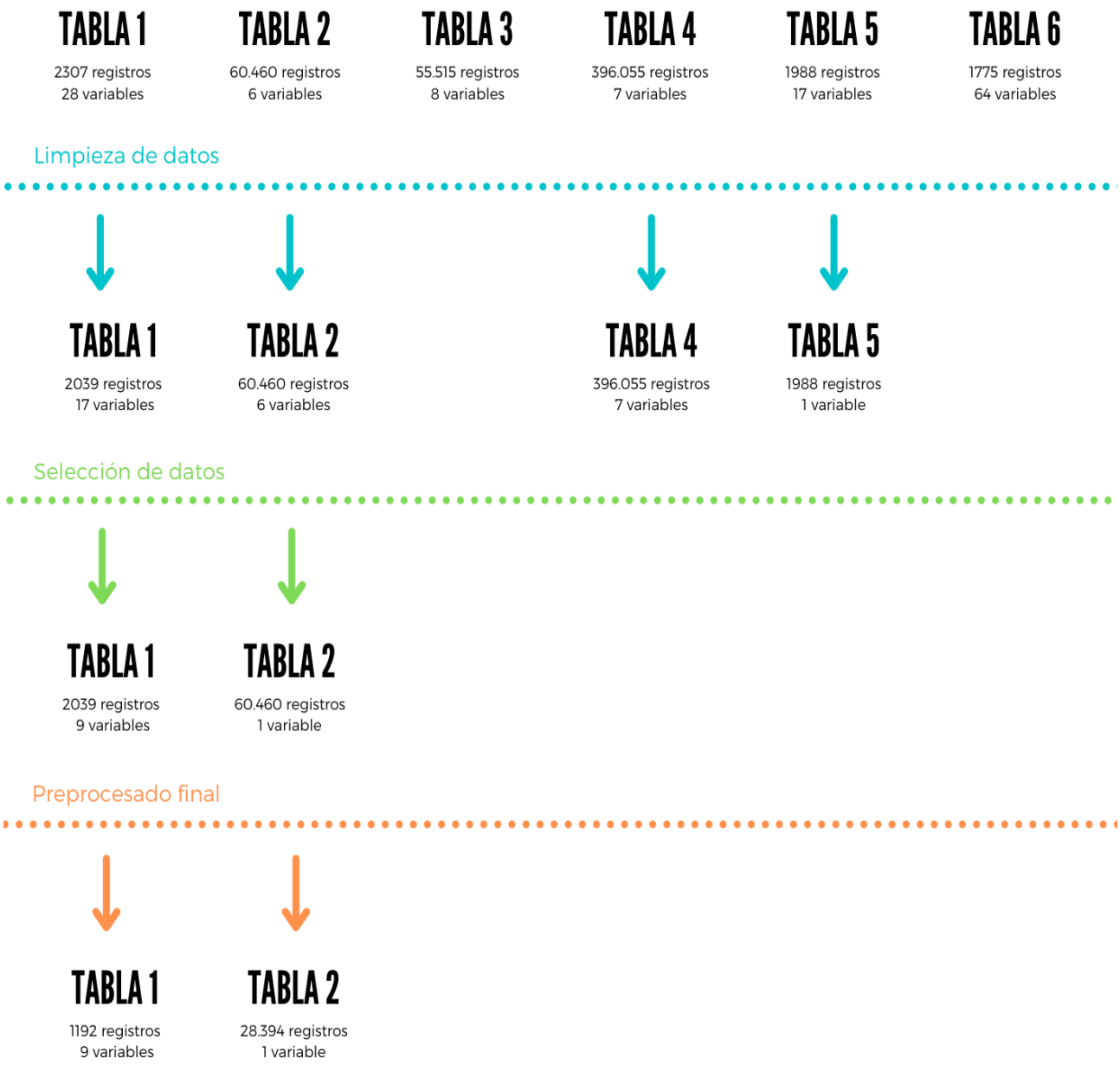

***Figura 11**. Diagrama de flujo para la selección de tablas, variables y registros.*

- **Diagnóstico de Urgencias**

De los 1194 casos, más de la mitad (un 57,13%) llegaron a Urgencias con motivo de dificultad respiratoria. En orden decreciente, los siguientes diagnósticos más comunes fueron fiebre (12,58%), cuadro catarral (8,72%), tos (7,05%), malestar general (3,44%), deterioro de paciente oncológico (2,77%), diarrea (1,17%) y malestar (1,01%). Por debajo del 1% se encuentran, entre otros, los diagnósticos disuria, desorientación, mareo, náuseas o vómitos, astenia, dificultad en diuresis, dolor lumbar, dolor en el pecho/dolor torácico, dolor al tragar, síncope y dolor costal.

Tabla 8. *Métricas de estadística descriptiva de las variables numéricas recogidas en la Tabla 1ª.*

	Media	**Mediana**	**Moda**	**Desviación estándar**	**Rango**
Edad (años)	68	69	64 y 71	15,8	[21 - 106]
Temperatura (ºC)	36,76	36,6	36,5	0,83	[33,2 - 39,8]
FC (lpm)	90,02	89	85	16,58	[41 - 170]
PaO_2 (mmHg)	92,21	94	95	6,84	[44 - 99]
PA máxima (mmHg)	131,74	131	130	21,02	[60 - 198]
PA mínima (mmHg)	75,20	76	70	12,74	[31 - 127]

Se observa a primera vista en los datos numéricos una edad avanzada en los pacientes, temperaturas y FC moderadas y PaO_2 relativamente elevadas. Los datos medios de PA sugieren que el paciente promedio es una persona con la tensión elevada, probablemente con una hipertensión tipo 2.

Analizando más detalladamente los datos de PA mínima y máxima, se puede hacer una clasificación de los pacientes en categorías tal como se expone en la Tabla 6. De estos, solamente un 26,59% se encontraban dentro de una categoría normal. Un 12,5% presentaban tensión elevada, un 21,32% hipertensión nivel 1, un 38,42% hipertensión nivel 2 y solamente un 1,17% presentaba hipertensión de nivel 3.

Llegados a este punto, se normalizaron los datos y se obtuvo un mapa de correlaciones, donde la defunción del paciente se seleccionó como variable dependiente. A fin de facilitar la visualización de los datos y su trabajo posterior, se agruparon los datos de "Motivo de alta" en 2 variables: defunción, referente a "fallecimiento" y supervivencia, donde se incluyeron "alta voluntaria", "domicilio", "traslado al hospital" y "traslado a un centro sociosanitario".

Los resultados, mostrados en la Figura 12 de la siguiente página, arrojan información relevante para conocer los principales motivos de fallecimiento en función de las variables de la Tabla 1ª. Se observa la edad como principal factor de riesgo, y se aprecia una tenue influencia de la PA y la PaO_2. A grandes rasgos, la temperatura y FC medidas en Urgencias no parecen tener mayor importancia. El diagnóstico que sobresale en el histograma, referido a "Dificultad respiratoria" destaca también.

Por otra parte, así como las estadísticas muestran un mayor riesgo para los hombres (véase el Capítulo 2) y aunque en este estudio el porcentaje de fallecidos relativo a varones es superior, no deben hacerse conjeturas, pues los registros de hombres en la base de datos superan significativamente a los de mujeres.

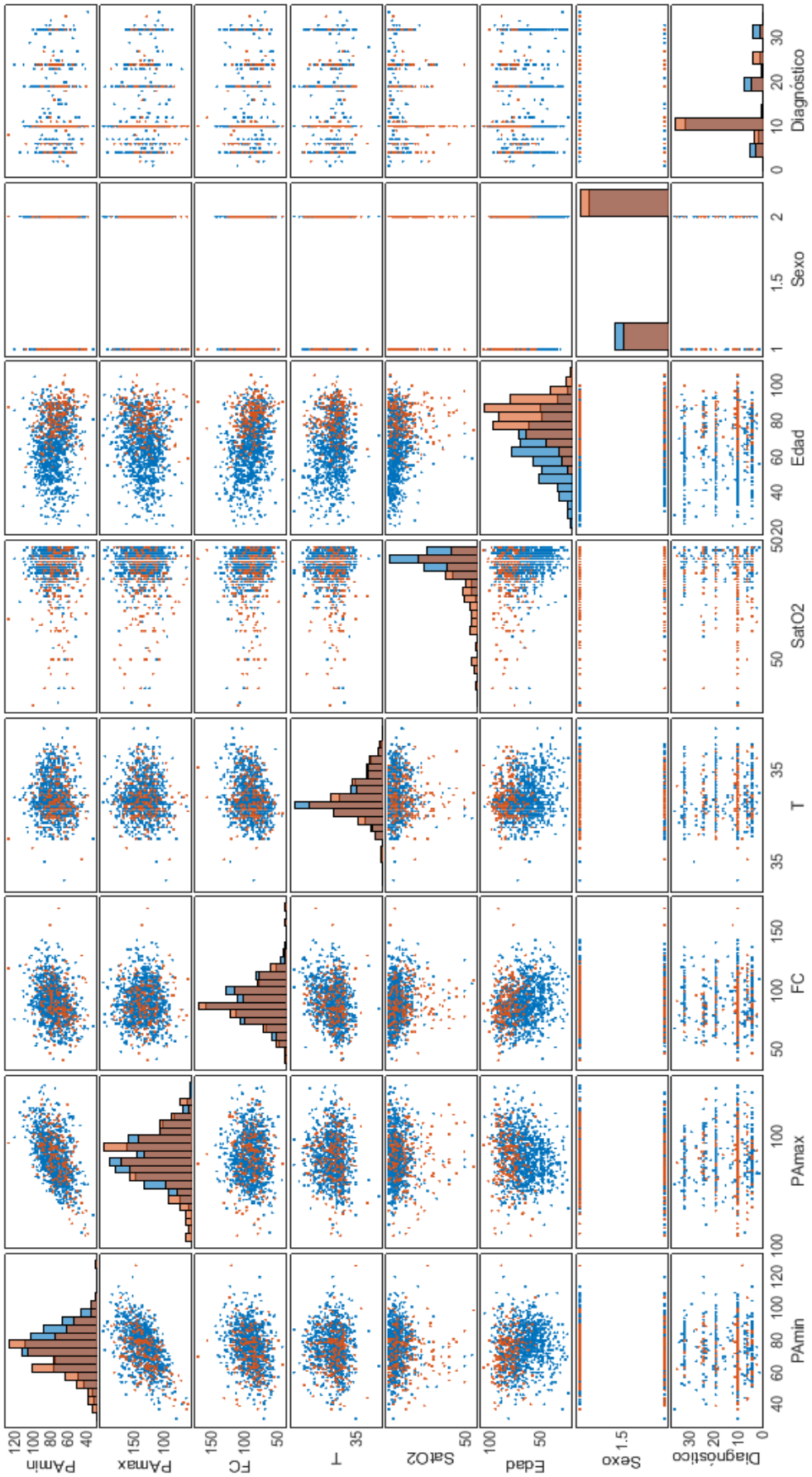

Figura 12. *Mapa de correlaciones entre variables de la Tabla 1ª. En rojo se muestran los pacientes que fallecieron y en azul los que sobrevivieron.*

5.1.4.2. Tabla 2ª

En esta tabla se recogen los distintos medicamentos pautados a cada paciente durante su ingreso. Con propósito de incluirlos como un indicador que ofreciese mayor claridad al clasificador, se analizaron por separado las medicaciones pautadas al total de individuos y a los individuos que fallecieron (5918 registros).

En términos generales, a los pacientes se les pautó con una frecuencia mayor fármacos agrupados dentro de electrolitos, paracetamol, etanol, bemiparina, clorhexidina, omeprazol, hidroxicloroquina, oxígeno, cloruro de sodio, azitromicina, metilprednisolona, dipirona, ritonavir y lopinavir, ceftriaxona, glucosa y electrolitos con carbohidratos. Atendiendo al código ATC, estos medicamentos son mayoritariamente de acción sobre la sangre y órganos hematopoyéticos (generalmente antitrombóticos, sustitutos de plasma y soluciones para infusión) y antiinfecciosos de uso sistémico (antibacterianos y antivirales) [47].

Respecto a los pacientes que fallecieron, se observó la misma pauta de fármacos con ligeras modificaciones. De más a menos recetados, los fármacos que se pautaron a los pacientes fallecidos fueron electrolitos, paracetamol, etanol, bemiparina, clorhexidina, omeprazol, cloruro de sodio, oxígeno, metilprednisolona, midazolam, hidroxicloroquina, glucosa, morfina, ceftriaxona, electrolitos con carbohidratos, dipirona y azitromicina.

Teniendo en cuenta que la morfina y el midazolam son fármacos sedantes paliativos, no se puede contemplar ninguna diferencia significativa entre el subgrupo que falleció y el total de pacientes en cuanto a su pauta de medicamentos.

5.2. CONSTRUCCIÓN DE UN MODELO DE APRENDIZAJE AUTOMÁTICO

Considerando los datos resultantes del análisis estadístico, se decide tomar como variables independientes para el algoritmo de predicción de evolución de pacientes hospitalizados por COVID-19 las variables que se recogen en la Tabla 9.

***Tabla 9**. Variables seleccionadas para el desarrollo del predictor.*

Algoritmo de clasificación	
Variable dependiente (a predecir)	**Variables independientes (predictoras)**
Fallecimiento (Sí/No)	Tabla 1ª: - Edad - Diagnóstico - PA mínima - PA máxima - PaO_2

Se toman estos datos y se entrenan los distintos tipos de clasificadores en la herramienta de MATLAB. Los distintos clasificadores se entrenan tanto para una validación cruzada *k-fold* con 5 particiones del subconjunto (*5-fold Cross-Validation*), como para una validación tipo *hold-out* con un 20% de los datos reservados para validación (*Hold-out Validation 20%*). Dado que las características del conjunto de datos no permitían entrenar la totalidad de los clasificadores, se entrenaron los 15 modelos disponibles para el *dataset* reducido.

- ***5-fold Cross-Validation***

El método con mejores resultados fue SVM *Medium Gaussian*, con un 86% de precisión y un núcleo (*kernel*) gaussiano de escala 2,4. La matriz de confusión binaria asociada a este clasificador y su curva ROC se observan en la Figura 13 y la Figura 14.

Sin embargo, es fácil ver que la precisión del clasificador viene dada por la facilidad del mismo a la hora de detectar cuándo un paciente no tiene riesgo de fallecer por COVID-19: su tasa de predicción de fallecimiento es baja y su tasa de falsos negativos muy elevada.

Se investigó de este modo entre los demás clasificadores entrenados y se seleccionó el clasificador *RUSBoosted Trees*, con un 74,8% de valor de predicción (matriz de confusión y curva ROC disponibles en la Figura 15 y la Figura 16).

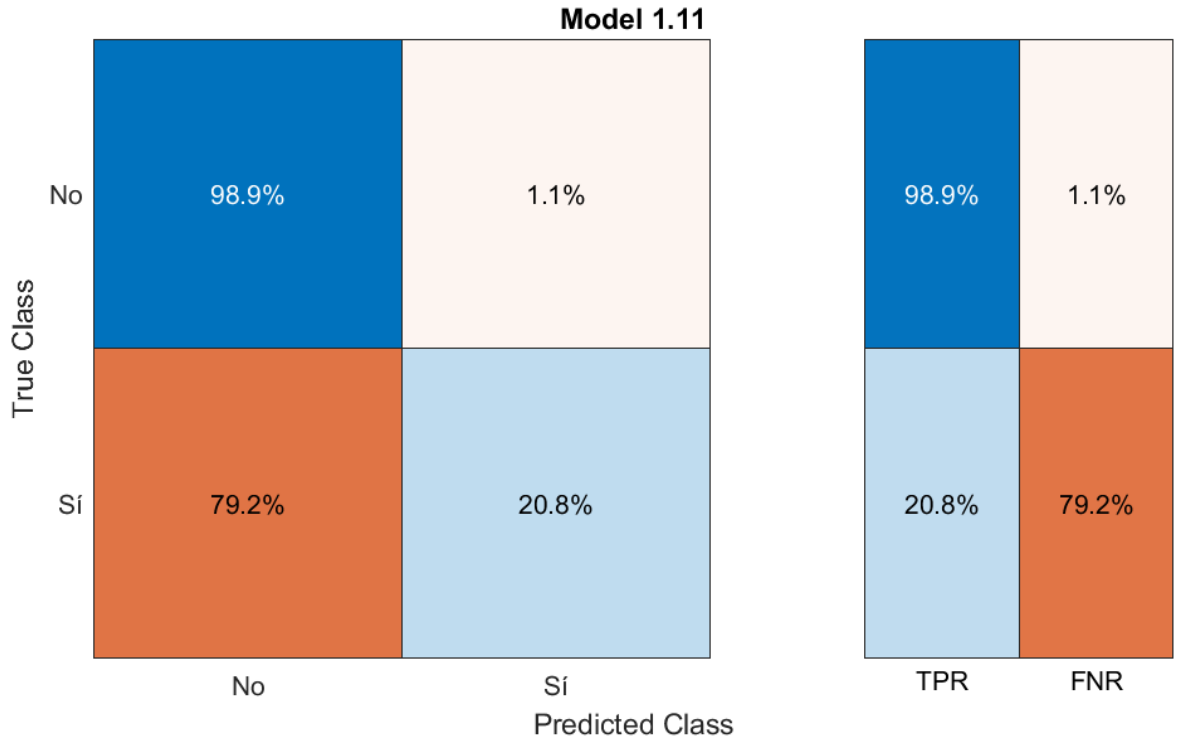

***Figura 13**. Matriz de confusión resultante del método SVM Medium Gaussian con validación cruzada 5-fold.*

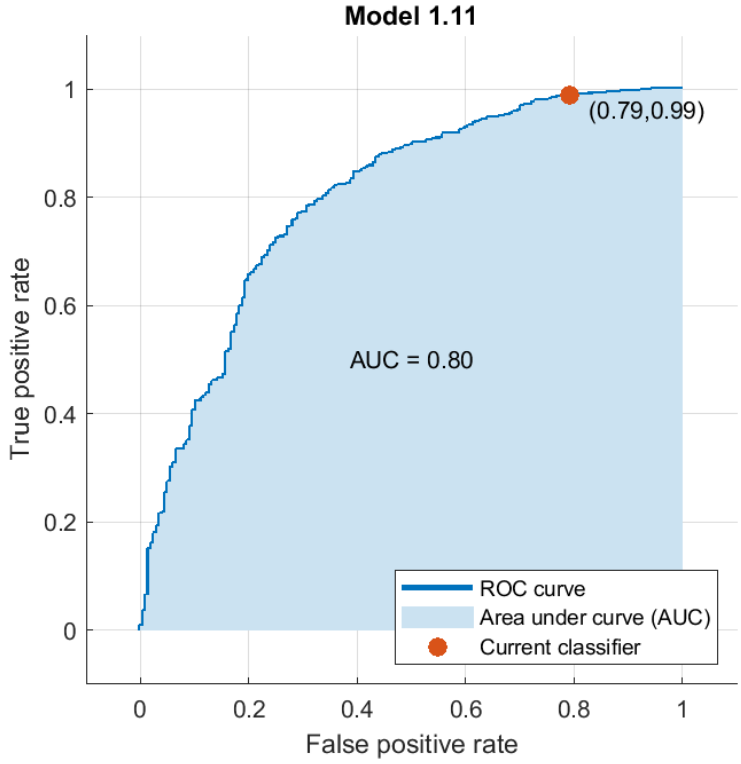

Figura 14. *Curva ROC resultante del método SVM Medium Gaussian con validación cruzada 5-fold.*

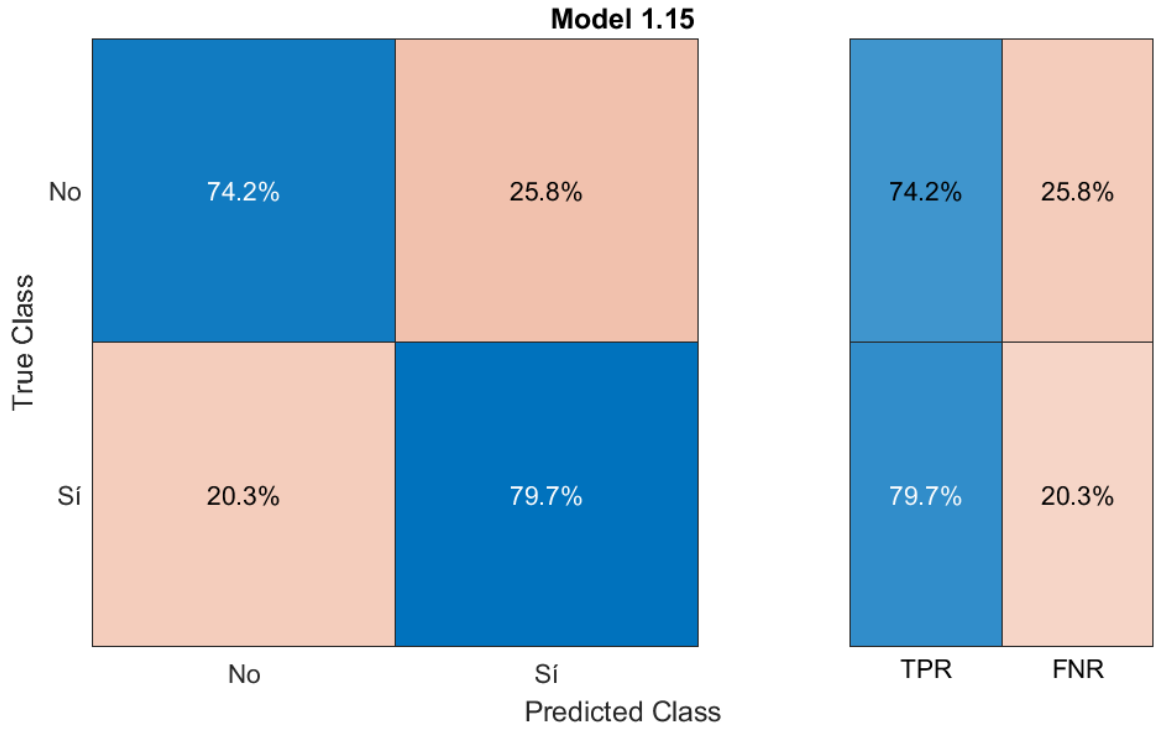

Figura 15*. Matriz de confusión resultante del método RUSBoosted Trees con validación cruzada 5-fold.*

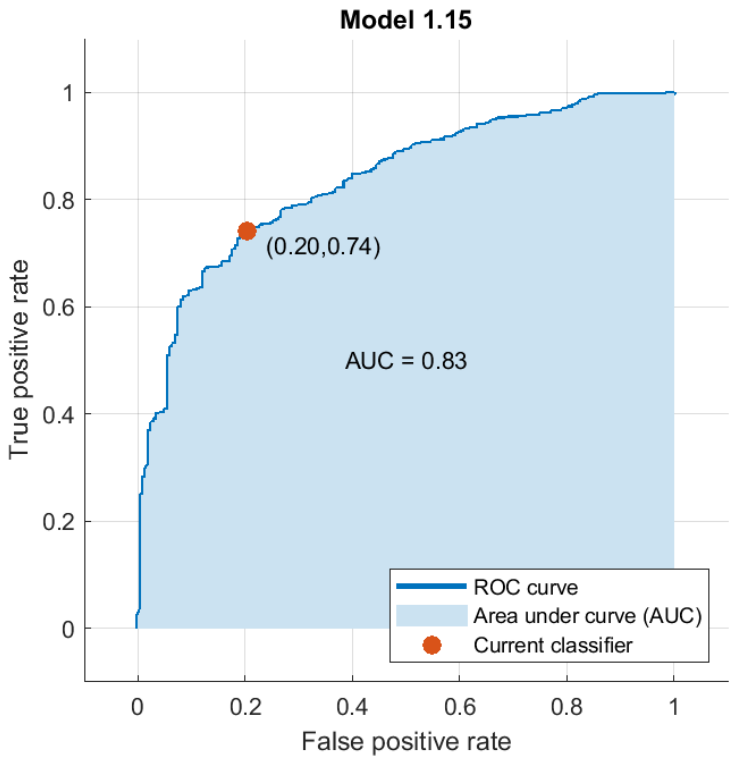

***Figura 16**. Curva ROC resultante del método RUSBoosted Trees con validación cruzada 5-fold.*

Se observa de este predictor que, aunque posee un valor inferior de precisión, su capacidad predictiva a la hora de determinar si un paciente tiene riesgo de fallecer es mucho más significativa. El área bajo la curva es ligeramente superior y su tasa de falsos negativos es muy inferior, sin sacrificar excesivamente el valor de la tasa de falsos positivos.

- ***Hold-out Validation 20%***

Para la segunda validación, el predictor con mejores resultados a priori fue el modelo *Kernel Naive Bayes*, con un 87% de precisión. Sin embargo, tal y como se observa en la Figura 17 y la Figura 18, sucede el mismo problema que en el caso de la validación cruzada *k-fold*.

Así pues, se examinaron de la misma forma que anteriormente el resto de los modelos en busca de mejores resultados. De nuevo, se optó por el modelo *RUSBoosted Trees*, con un 74,4% de precisión. En la Figura 19 y la Figura 20 pueden observarse su matriz de confusión y su curva ROC. Se observa en estas figuras una mejoría notable en la capacidad predictiva del modelo.

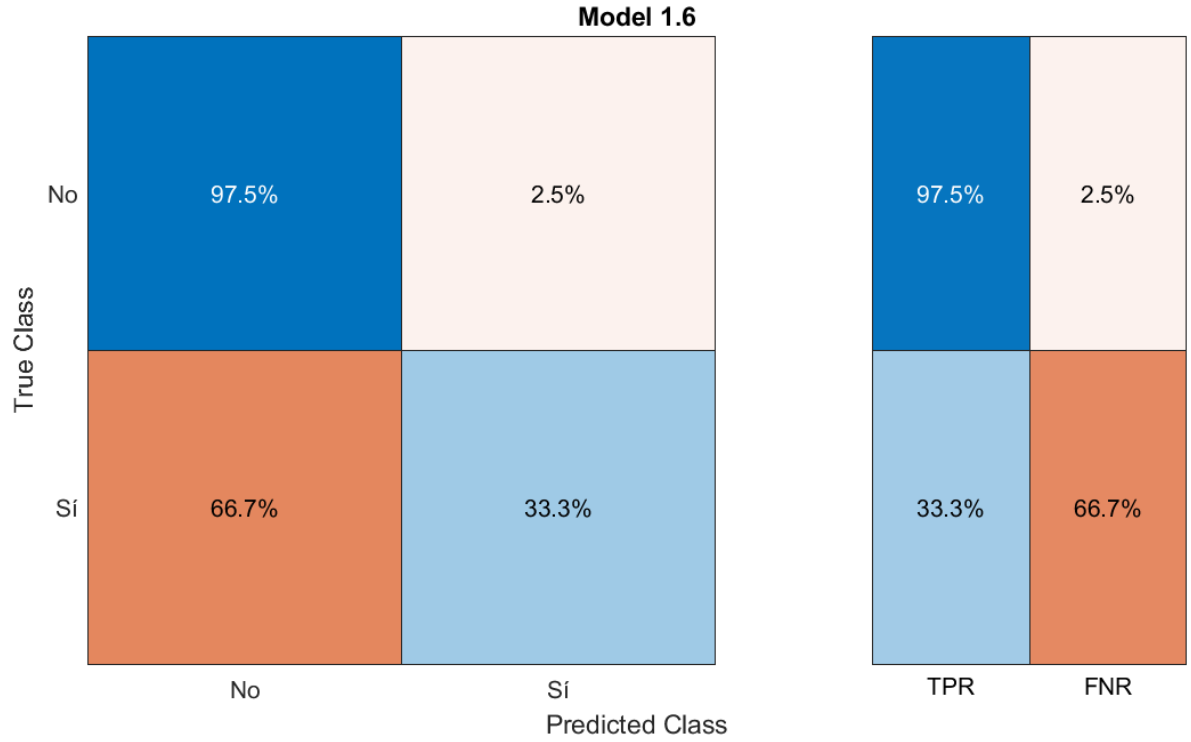

***Figura 17**. Matriz de confusión resultante del método Kernel Naive Bayes con validación hold-out (20%).*

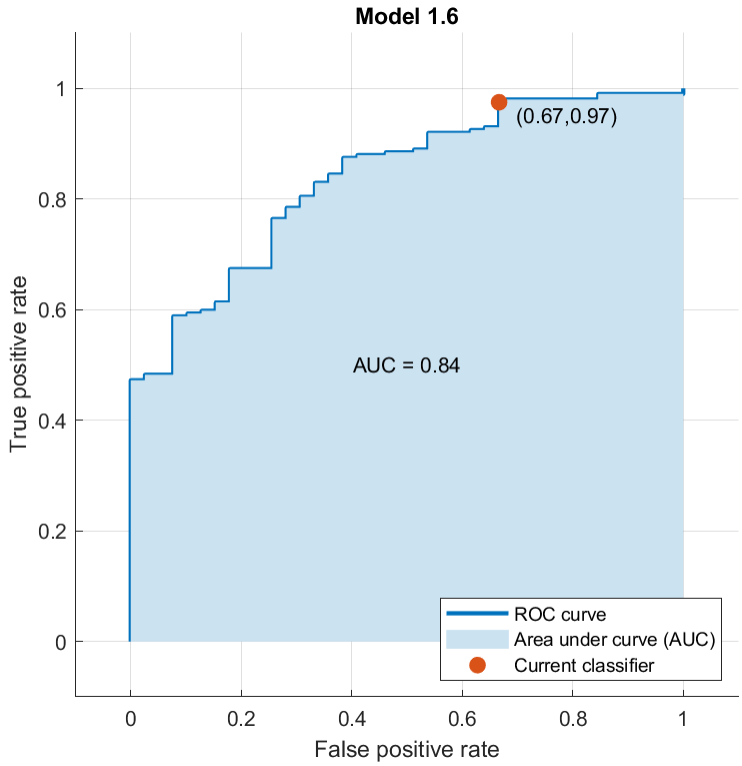

***Figura 18**. Curva ROC resultante del método Kernel Naive Bayes con validación hold-out (20%).*

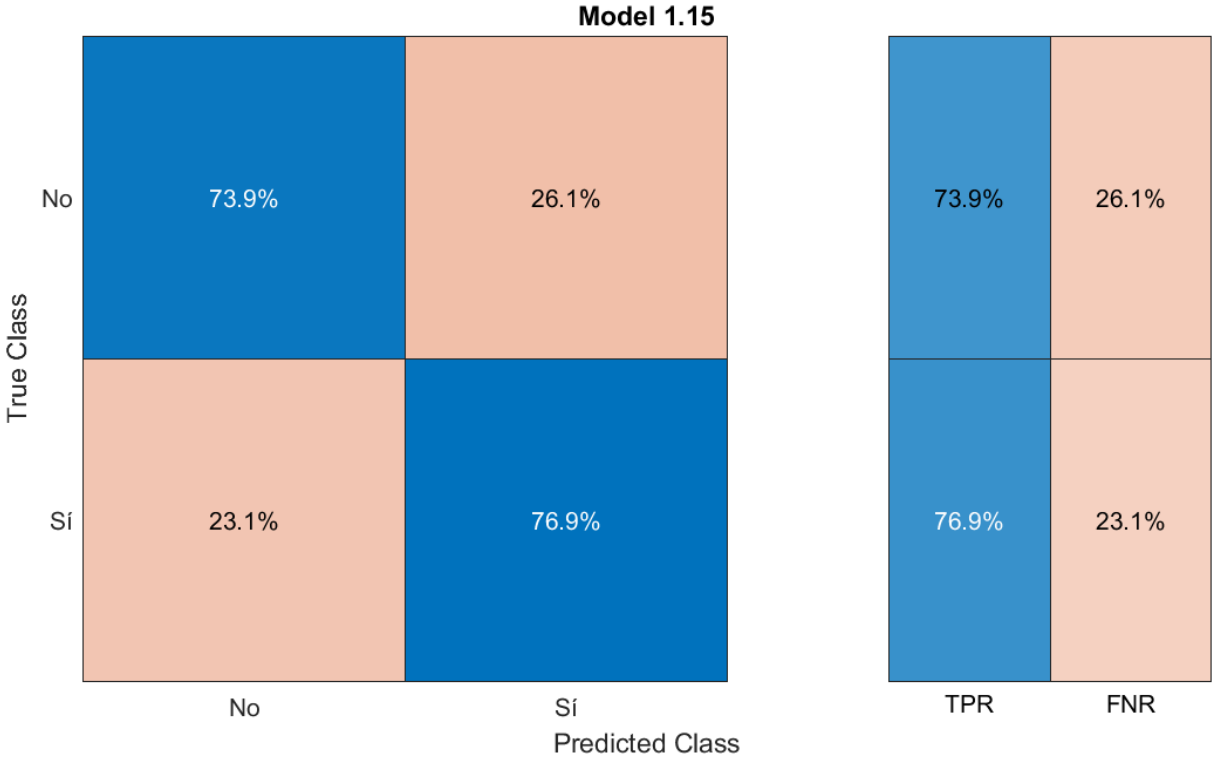

Figura 19. *Matriz de confusión resultante del método RUSBoosted Trees con validación hold-out (20%).*

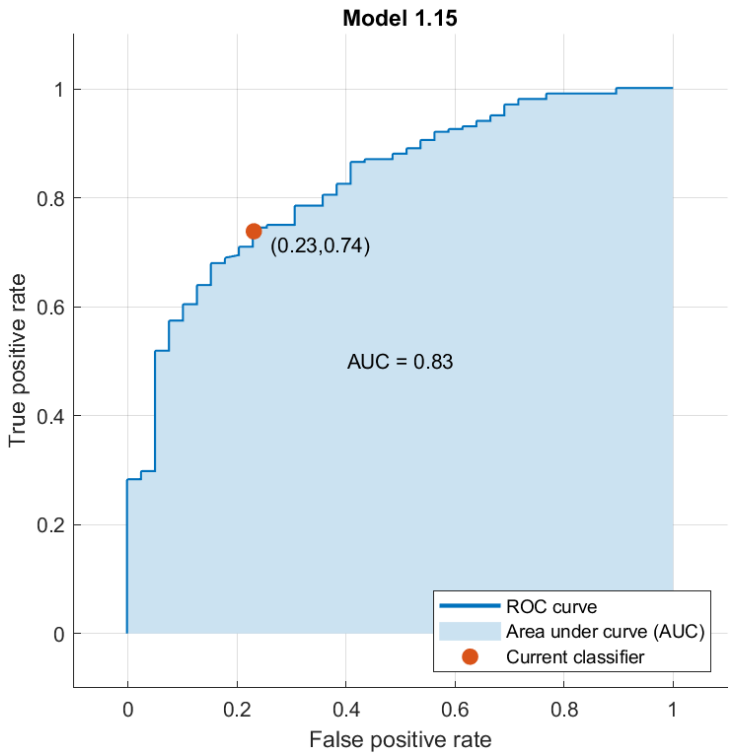

Figura 20*. Curva ROC resultante del método RUSBoosted Trees con validación hold-out (20%).*

Finalmente se seleccionó entre los dos modelos sobresalientes en cada tipo de validación, considerando su capacidad predictiva en cuanto a predicción de pacientes con alta probabilidad de fallecer. Se trataba de dos algoritmos muy similares, con situaciones en sus curvas ROC casi idénticas y precisiones parecidas. Teniendo esto en cuenta, se seleccionó el primer modelo, dado que se observaba una mejor capacidad para discernir en aquellos casos en que el paciente presentaba un alto riesgo de defunción.

Así pues, el método seleccionado para predecir la evolución de pacientes hospitalizados por COVID-19 fue un modelo *RUSBoosted Trees* que utiliza técnicas de validación cruzada *k-fold* con 5 particiones del subconjunto.

En la Figura 21 se observan las aproximaciones realizadas por el algoritmo en cada variable para predecir las etiquetas del conjunto original de entrada. Se aprecia la influencia de la edad, de una saturación de oxígeno baja y la relevancia de un cuadro diagnóstico en Urgencias de dificultad respiratoria. Si bien no es fácil interpretar los datos relativos a la PA, sin estos la precisión del clasificador disminuye.

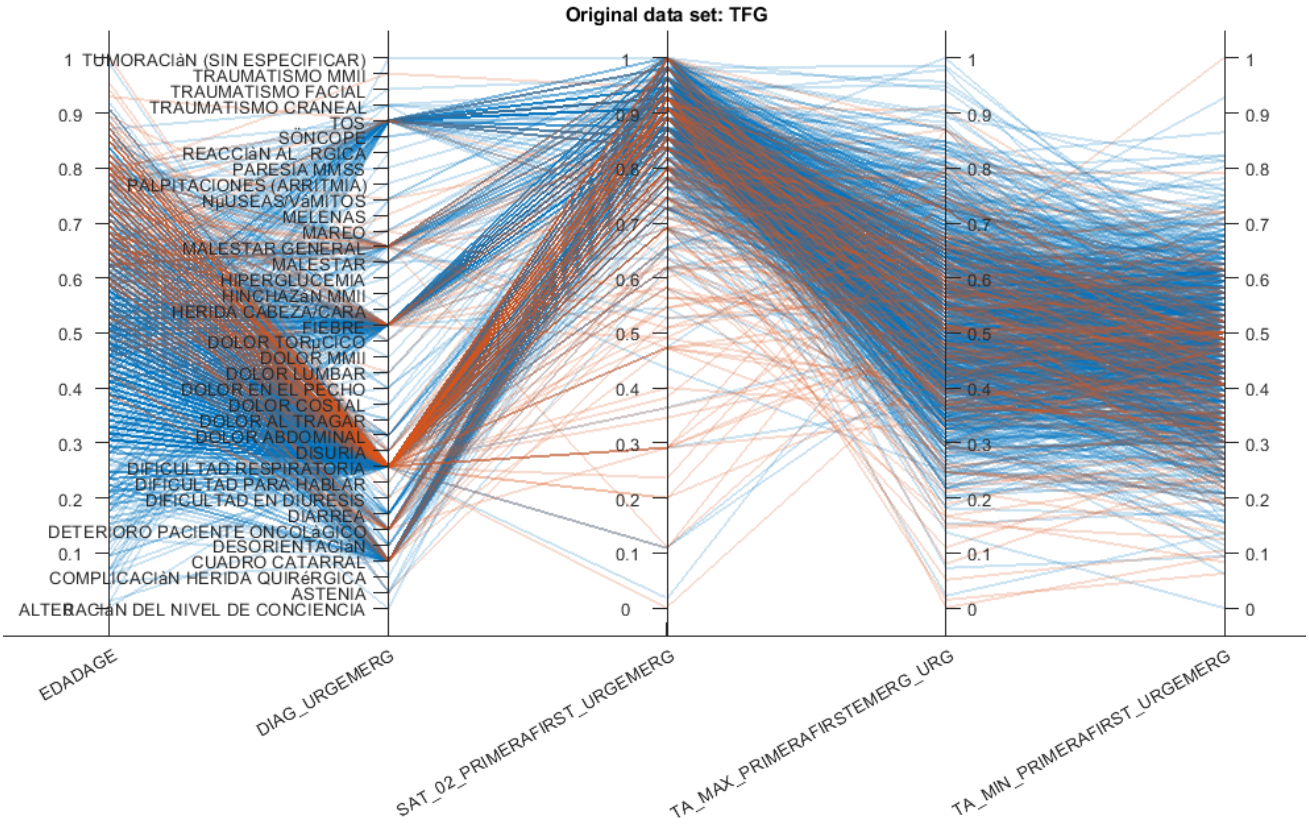

Figura 21. *Gráfico de coordenadas paralelas del clasificador utilizado para la predicción. El color rojo muestra a los pacientes fallecidos y el azul a los que sobrevivieron.*

Por otra parte, estudiando el comportamiento del algoritmo a lo largo de varias iteraciones, puede observarse que se mantiene su capacidad predictora, según se observa en la Figura 22 para 10 iteraciones distintas.

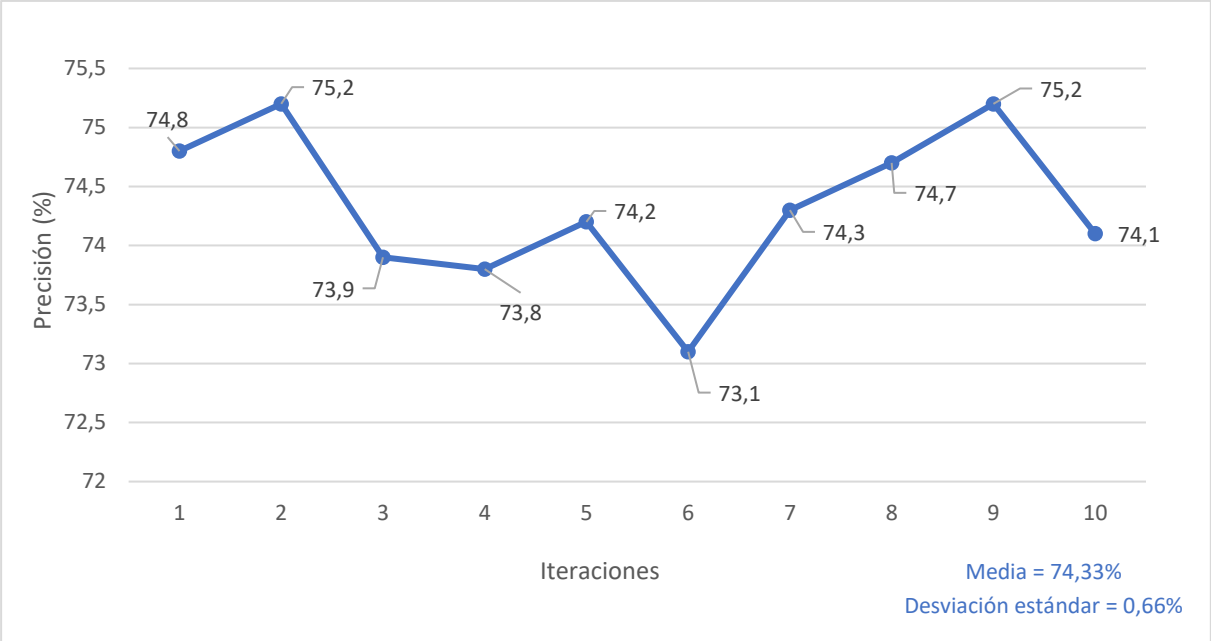

***Figura 22**. Precisión del clasificador RUSBoosted Trees a lo largo de 10 iteraciones.*

6. CONCLUSIONES

En situaciones de incertidumbre y de colapso en centros sanitarios, disponer de sistemas o programas que proporcionen ayuda al diagnóstico es clave para conseguir llevar adelante la situación. Fijando esta idea como objetivo, este trabajo se presenta como un ejemplo de desarrollo de un algoritmo de clasificación con aplicación en el ámbito médico, concretamente para centros sanitarios donde exista un gran volumen de pacientes contagiados por COVID-19.

Para este tipo de situaciones, hacer un diagnóstico veraz en la llegada de los pacientes a Urgencias es fundamental para asegurar su buena evolución. Así, el algoritmo desarrollado se basa esencialmente en las constantes tomadas a los pacientes en Urgencias, de forma que sus resultados puedan influir directamente en el triaje y tratamiento de los pacientes.

Del análisis de los datos llevado a cabo en el Capítulo 5 se observan conjuntos de datos coherentes con la realidad clínica y con la sintomatología propia de la COVID-19 (véase el Capítulo 1). Sin embargo, durante la fase previa de limpieza de los datos se encontraron severas limitaciones que probablemente habrán limitado el funcionamiento del clasificador.

En primer lugar, la existencia de datos copiados en distintas variables genera desconfianza en la validez de los datos y obliga a quien trabaja con ellos a hacer asunciones que pueden no ser correctas. De hecho, para el clasificador desarrollado se ha asumido que los primeros datos de constantes tomados a la urgencia (idénticos a los datos de las últimas constantes) son correctos, a pesar de no haber ninguna garantía de ello.

En segundo lugar, sin comprometer la anonimidad de los pacientes, hubiese sido de gran ayuda incluir antecedentes clínicos como obesidad, enfermedades crónicas o tabaquismo asociados a su ID. De este modo, se podría haber incluido la existencia de este factor como una variable de alta capacidad predictiva o haberse trabajado con los datos de la medicación pautada (Tabla 2ª) con mayor conocimiento. No obstante, según el análisis de los fármacos en el ingreso llevado a cabo en el Capítulo 5, no se aprecian diferencias significativas en el tratamiento de los pacientes que fallecieron y en el de los que sobrevivieron.

En tercer lugar, a pesar del gran aspecto positivo que supone la cantidad de datos completos ofrecido por HM Hospitales, dado que el clasificador objetivo se rige por la separación entre el fallecimiento o supervivencia de los pacientes, la baja cantidad de pacientes fallecidos en el *dataset* dificulta el aprendizaje. Por otra parte, es una limitación el hecho de no disponer de datos externos para realizar una validación.

Finalmente, existen limitaciones adicionales asociadas al uso del software y a la selección de las variables. Un software y un ordenador más potentes son imprescindibles para un procesamiento más profundo de los datos; y probablemente un estudio más trascendente de estos sea necesario para identificar parámetros que ofrezcan unos mejores resultados. Sin embargo, aunque algunos parámetros como los resultados de laboratorio han demostrado ser realmente importantes para

determinar la gravedad de pacientes con COVID-19, incluir este tipo de parámetros en un clasificador requiere que a todos los pacientes se les realicen estas pruebas, algo que no siempre es posible.

A pesar de las limitaciones, el algoritmo desarrollado ha resultado ser significativamente preciso. Se trata de un algoritmo basado en un clasificador *RUSBoosted Trees*, un tipo de clasificador de métodos conjuntos. Se ha utilizado el método de validación cruzada *k-fold* con 5 particiones de datos para su entrenamiento. Utiliza 5 variables predictoras, la edad, la PaO_2, las PA máxima y mínima y el diagnóstico a la urgencia. Y, si bien es susceptible a mejoras, desde un punto de vista holístico y suponiendo que las asunciones llevadas a cabo son correctas, los resultados se consideran satisfactorios.

Dentro de todas las limitaciones mencionadas, el predictor ofrece un 74,8% de precisión y bajas tasas de error, por lo que podría significar una buena herramienta de ayuda para el predecir la gravedad de pacientes que llegan con un cuadro clínico de COVID-19 a Urgencias. El predictor asegura un pronóstico correcto a al menos un 79,7% de los pacientes que lleguen con un elevado riesgo de morir.

El hecho de que este porcentaje sea inferior al de proyectos paralelos como los mencionados en el Capítulo 1 se atribuye generalmente a las limitaciones descritas y a la necesidad de un proyecto de investigación más exhaustivo que el que se puede abarcar en este trabajo de final de grado.

Finalmente, aunque escapa a los objetivos del proyecto entrar en más detalle, cabe mencionar que el fin de este tipo de algoritmos es siempre el de conformar parte de un sistema de apoyo a la práctica clínica, no pudiendo imponerse nunca el resultado de estos a una opinión médica fundamentada.

7. REFERENCIAS

[1] K. H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," Nature Publishing Group, Oct. 2018. doi: 10.1038/s41551-018-0305-z.

[2] Naciones Unidas, "Cronología de la pandemia del coronavirus y la actuación de la Organización Mundial de la Salud | Noticias ONU." https://news.un.org/es/story/2020/04/1472862 (accessed Jun. 19, 2021).

[3] R. Mojica-Crespo and M. M. Morales-Crespo, "Pandemia COVID-19, la nueva emergencia sanitaria de preocupación internacional: una revisión," *Medicina de Familia. SEMERGEN*, vol. 46, Aug. 2020, doi: 10.1016/j.semerg.2020.05.010.

[4] Organización Mundial de la Salud (OMS), "Cronología de la respuesta de la OMS a la COVID-19." https://www.who.int/es/news/item/29-06-2020-covidtimeline (accessed Jun. 19, 2021).

[5] C. y B. S.-G. de P. Ministerio de Sanidad, "Notas de Prensa - La EMA recomienda la autorización de la primera vacuna frente a la COVID-19." https://www.mscbs.gob.es/gabinete/notasPrensa.do?metodo=detalle&id=5177 (accessed Jun. 20, 2021).

[6] ISCII CNE - RENAVE, "COVID-19 | Evolución pandemia." https://cnecovid.isciii.es/covid19/ (accessed Jun. 20, 2021).

[7] G. de E. Ministerio de Sanidad, "Informe de actividad del proceso de vacunación | GIV COVID-19," Jun. 18, 2021. https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/documentos/Informe_GIV_comunicacion_20210618.pdf (accessed Jun. 20, 2021).

[8] H. Ritchie *et al.*, "Coronavirus Pandemic (COVID-19) – the data | Statistics and Research - Our World in Data." https://ourworldindata.org/coronavirus-data (accessed Jun. 21, 2021).

[9] Y. Junejo *et al.*, "Novel SARS-CoV-2/COVID-19: Origin, pathogenesis, genes and genetic variations, immune responses and phylogenetic analysis," *Gene Reports*, vol. 20, Sep. 2020, doi: 10.1016/j.genrep.2020.100752.

[10] B. Hu, H. Guo, P. Zhou, and Z. L. Shi, "Characteristics of SARS-CoV-2 and COVID-19," *Nature Reviews Microbiology*, vol. 19, no. 3. Nature Research, pp. 141–154, Mar. 01, 2021, doi: 10.1038/s41579-020-00459-7.

[11] M. Palacios Cruz, E. Santos, M. A. Velázquez Cervantes, and M. León Juárez, "COVID-19, a worldwide public health emergency," *Revista Clinica Espanola*, vol. 221, no. 1. Elsevier Doyma, pp. 55–61, Jan. 01, 2021, doi: 10.1016/j.rce.2020.03.001.

[12] Redacción Médica, "Foto del coronavirus: así es de verdad el SARS-CoV-2."

[13] E. Estrada, "Fractional diffusion on the human proteome as an alternative to the multi-organ damage of SARS-CoV-2," *Chaos*, vol. 30, no. 8, Aug. 2020, doi: 10.1063/5.0015626.

[14] GRUPO DE ANALISIS CIENTÍFICO DE CORONAVIRUS DEL ISCIII, "Informe del GACC - ISCII | Factores de riesgo en la enfermedad por SARS-CoV-2 (COVID-19)." Accessed: Jun. 21, 2021. [Online].

[15] Centers for Disease Control and Prevention, "Scientific Evidence for Conditions that Increase Risk of Severe Illness | COVID-19 ." https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/underlying-evidence-table.html (accessed Jun. 21, 2021).

[16] Organización Mundial de la Salud, "Información basíca sobre la COVID-19." https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19 (accessed Jun. 21, 2021).

[17] F. y C. (SEFAC) Sociedad Española de Farmacia Clínica, "Evolución de la COVID-19 en el paciente: fases y características." https://www.sefac.org/para-profesionales-publicaciones-sefac-publicaciones-sefac-materiales-disponibles-para-el-socio-5 (accessed Jun. 25, 2021).

[18] P. Joshi, *Artificial Intelligence with Python*. Birminhham, UK, 2017.

[19] R. E. López Briega, *Libro online de la comunidad de Inteligencia Artificial de Argentina*. .

[20] A. K. Triantafyllidis and A. Tsanas, "Applications of Machine Learning in Real-Life Digital Health Interventions: Review of the Literature," *Journal of Medical Internet Research*, vol. 21, no. 4, Apr. 2019, doi: 10.2196/12286.

[21] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. 2011.

[22] J. V. Manjón Herrera, "Tema 6. Clasificación de patrones. [Diapositivas de PowerPoint] - Asignatura Imágenes Biomédicas. Universitat Politècnica de València," 2020. Accessed: Jun. 23, 2021. [Online]. Available: https://poliformat.upv.es/access/content/group/GRA_13062_2020/Tema6.Analisis_parte2_.pdf.

[23] A. Zheng, *Evaluating Machine Learning Models*, 1st ed. United States of America: O'Reilly Media, Inc., 2015.

[24] L. Pérez-Planells, J. Delegido, J. P. Rivera-Caicedo, and J. Verrelst, "Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos," *Revista de Teledeteccion*, vol. 2015, no. 44, pp. 55–65, Dec. 2015, doi: 10.4995/raet.2015.4153.

[25] G. Pérez, "Peligros del uso de los big data en la investigación en salud pública y en epidemiología," *Gaceta Sanitaria*, vol. 30, no. 1, pp. 66–68, Jan. 2016, doi: 10.1016/j.gaceta.2015.09.007.

[26] World Health Organization, "National EHR system exists | European Health Information Gateway." https://gateway.euro.who.int/en/indicators/ehealth_survey_84-has-a-national-ehr-system/visualizations/#id=31759&tab=table (accessed Jun. 25, 2021).

[27] S. J. Mooney and V. Pejaver, "Big Data in Public Health: Terminology, Machine Learning, and Privacy," *Annual review of public health*, vol. 39, p. 95, Apr. 2018, doi: 10.1146/ANNUREV-PUBLHEALTH-040617-014208.

[28] Instituto Europeo de Salud y Bienestar, "El desafío del Big Data en los Sistemas de Salud ." https://institutoeuropeo.es/articulos/insights/el-desafio-del-big-data-en-los-sistemas-de-salud/ (accessed Jun. 22, 2021).

[29] HM Hospitales, "COMUNICADO: COVID DATA SAVE LIVES." https://www.hmhospitales.com/prensa/notas-de-prensa/comunicado-covid-data-save-lives (accessed Jun. 22, 2021).

[30] HM Hospitales, "Covid Data Save Lives." https://www.hmhospitales.com/coronavirus/covid-data-save-lives (accessed Jun. 22, 2021).

[31] Hospital Universitario 12 de Octubre, "El 12 de Octubre crea una calculadora que predice el riesgo de mortalidad por COVID-19 ," Apr. 19, 2021. https://www.comunidad.madrid/hospital/12octubre/noticia/12-octubre-crea-calculadora-predice-riesgo-mortalidad-covid-19 (accessed Jun. 28, 2021).

[32] "COR+12: imas12 Mortality Score for COVID-19," *Hospital 12 de Octubre*. https://utrero-rico.shinyapps.io/COR12_Score/ (accessed Jun. 23, 2021).

[33] Quirónsalud, "Big data para pronosticar la evolución del paciente Covid ," May 24, 2021. https://www.tucanaldesalud.es/es/tecnologia/articulos/big-data-pronosticar-evolucion-paciente-covid (accessed Jun. 28, 2021).

[34] J. Torres-Macho *et al.*, "The PANDEMYC Score. An Easily Applicable and Interpretable Model for Predicting Mortality Associated With COVID-19," *Journal of Clinical Medicine*, vol. 9, no. 10, p. 3066, Sep. 2020, doi: 10.3390/jcm9103066.

[35] PERMANENTE MEDICINE, "Nueva herramienta de evaluación ayuda a predecir los resultados de COVID-19 ," Nov. 20, 2020. https://permanente.org/new-assessment-tool-helps-predict-outcomes-for-covid-19/ (accessed Jun. 28, 2021).

[36] A. L. Sharp *et al.*, "Identifying patients with symptoms suspicious for COVID-19 at elevated risk of adverse events: The COVAS score," *American Journal of Emergency Medicine*, vol. 0, no. 0, 2020, doi: 10.1016/j.ajem.2020.10.068.

[37] HM Hospitales, "HM Hospitales y el Massachusetts Institute of Technology unen sinergias para predecir el comportamiento del Covid-19," Jul. 20, 2020. https://www.hmhospitales.com/prensa/notas-de-prensa/hm-hospitales-y-massachusetts-institute-of-technology-unen-sinergias-para-predecir-comportamiento-covid-19 (accessed Jun. 28, 2021).

[38] MathWorks, "MATLAB ." https://es.mathworks.com/products/matlab.html (accessed Jun. 28, 2021).

[39] MathWorks, "Choose Classifier Options." https://es.mathworks.com/help/stats/choose-a-classifier.html (accessed Jun. 29, 2021).

[40] Statistics Solutions, "Análisis discriminante." https://www.statisticssolutions.com/discriminant-analysis/ (accessed Jun. 29, 2021).

[41] Rachel Nall, "Temperatura corporal: Rangos normales en adultos y niños," *Medical News Today*, Jun. 09, 2020.

[42] "¿Cómo se produce la Fiebre? ," *Hospital Clínic Barcelona*. https://www.clinicbarcelona.org/asistencia/cuida-tu-salud/fiebre/causas (accessed Jul. 06, 2021).

[43] Edward R. and M. D. Laskowski, "Frecuencia cardíaca: ¿cuál es la normal? ," *Clínica Mayo*, Oct. 20, 2020. https://www.mayoclinic.org/es-es/healthy-lifestyle/fitness/expert-answers/heart-rate/faq-20057979 (accessed Jul. 06, 2021).

[44] P. Oliver *et al.*, "Estudio de la oxigenación e interpretación de la gasometría arterial. Revisión," 2014. Accessed: Jul. 06, 2021. [Online]. Available: https://www.seqc.es/download/doc/62/2845/951224035/858217/cms/estudio-de-la-oxigenacion-e-interpretacion-de-la-gasometria-arterial-revision-2014.pdf/.

[45] S. Daru, "Optimización de la Perfusión. Hiperoxia: La otra cara de la moneda," Accessed: Jul. 06, 2021. [Online].

[46] A. F. Rubio-Guerra and A. F. Rubio-Guerra, "Nuevas guías del American College of Cardiology/American Heart Association Hypertension para el tratamiento de la hipertensión," *Medicina interna de México*, vol. 34, no. 2, pp. 299–303, Mar. 2018, doi: 10.24245/MIM.V34I2.2015.

[47] Agencia Española de Medicamentos y Productos Sanitarios (AEMPS), "Listados de principios activos por grupos ATC* e incorporación del pictograma de la conducción," May 25, 2021. https://www.aemps.gob.es/cuidadania/medicamentos-y-conduccion/industria_etiquetado_conduccion_listadosprincipios/ (accessed Jul. 08, 2021).

PRESUPUESTO

Desarrollo de un algoritmo de aprendizaje automático para predicción de evolución de pacientes hospitalizados por COVID-19

Documento II

Belén Ballester Viciano

Grado en Ingeniería Biomédica

Curso académico 2020 - 2021

Índice de contenidos

1. OBJETIVO

El objetivo de este documento es el de realizar una valoración económica estimada del proyecto llevado a cabo, basado en el desarrollo de un algoritmo de aprendizaje automático para predecir la evolución y gravedad de pacientes hospitalizados por COVID-19.

Para el desarrollo de este presupuesto se utiliza el software Arquímedes, de CYPE Ingenieros S.A.

2. PRESUPUESTO

2.1. PRESUPUESTOS PARCIALES

El informe de presupuestos parciales se divide en tres secciones diferenciadas: costes de mano de obra, costes de maquinaria y costes de materiales.

2.1.1. Costes de Mano de Obra

Se describen en esta sección los costes de recursos humanos asociados al desarrollo del trabajo. Para ello, se realiza una estimación del tiempo empleado en su desarrollo y de las remuneraciones aproximadas de los distintos participantes en el proyecto.

En la Tabla 10 se desglosan los costes de mano de obra y se incluye toda esta información. Se considera la contribución de Dª. Valery Naranjo Ornedo, catedrática de Universidad y tutora del presente proyecto; de D. Jorge Igual García, profesor titular de Universidad y encargado de cotutorizar el proyecto y Dª. Belén Ballester Viciano, como estudiante del Grado en Ingeniería Biomédica y autora del proyecto.

***Tabla 10**. Desglose de los costes de mano de obra.*

Descripción	Uds.	Precio unitario (€/h)	Cantidad	Coste imputable (€)
Catedrática de Universidad	h	42,00	12	504,00
Profesor titular de Universidad	h	35,00	40	1400,00
Estudiante	h	13,00	300	3900,00
			Total mano de obra:	5804,00 €

2.1.2. Costes de Maquinaria

En esta sección se tienen en cuenta los costes de hardware y de software necesarios para llevar a cabo el proyecto.

- **Costes de hardware**

Todo el proyecto se ha realizado con un único ordenador personal. Si bien las imágenes de tórax proporcionadas por HM Hospitales a CVBLab se encuentran almacenadas en los servidores del grupo de investigación, estos servidores no se consideran por no haber sido utilizadas las imágenes para el desarrollo del presente proyecto.

El coste que ha supuesto el uso del equipo personal se desglosa en la Tabla 11, teniendo en cuenta su periodo de amortización, puesto que el dispositivo no se adquirió de forma exclusiva para la realización del proyecto.

- **Costes de software**

En el coste de software se incluyen las licencias de los distintos softwares utilizados. Tal y como se explica en el Capítulo 3, se utilizan para la visualización y procesado de datos el software MATLAB® v.R2020b de MathWorks y el software de hojas de cálculo de Microsoft, Excel.

Los costes de esta sección se recogen en la Tabla 12, donde de nuevo se tiene en cuenta los periodos de amortización.

*T**abla 11**. Desglose de los costes de maquinaria relativos al hardware.*

Descripción	Uds.	Precio unitario (€)	Cantidad	Periodo de amortización (meses)	Intervalo amortizado (meses)	Coste imputable (€)
HP Intel®Core™ i5-6200U	1	699,00	u	72	6	58,25
					Total hardware:	58,25 €

***Tabla 12**. Desglose de los costes de maquinaria relativos al software.*

Descripción	Uds.	Precio unitario (€)	Cantidad	Periodo de amortización (meses)	Intervalo amortizado (meses)	Coste imputable (€)
Licencia MATLAB R2020b	1	800,00	u	12	6	400,00
Licencia Microsoft 365 Personal	1	69,00	U	12	6	34,50
					Total software:	434,50 €

2.1.3. Costes de Materiales

El único material adicional empleado para el desarrollo del proyecto ha sido la base de datos facilitada por HM Hospitales a CVLab, la cual es completamente gratuita. De esta forma, el coste de los materiales se puede desestimar por ser igual a cero.

2.2. PRESUPUESTO TOTAL

Para el cómputo del presupuesto total del proyecto se obtiene en primer lugar el presupuesto de ejecución material. A este importe se le añaden los porcentajes de gastos generales (13%) y de beneficio industrial (6%), dando lugar al precio total bruto. A continuación, se añade a este precio el Impuesto sobre el Valor Añadido (IVA), correspondiente a un 21% de dicho valor y se obtiene el presupuesto de ejecución por contrata.

El valor del presupuesto de ejecución por contrata, reflejado junto al resto de cálculos en la Tabla 13, representa el presupuesto total que supondría la realización del presente proyecto.

Tabla 13. *Presupuesto total del proyecto.*

Descripción	Coste (€)
Costes de mano de obra	5804,00
Costes de maquinaria	492,75
- Costes de hardware	58,25
- Costes de software	434,50
Presupuesto de ejecución material	**6296,75**
Gastos generales (13%)	818,58
Beneficio industrial (6%)	377,81
Precio total bruto	**7493,14**
IVA (21%)	1573,56
Presupuesto de ejecución por contrata	**9066,70 €**

Así pues, el coste total de la realización del proyecto realizado asciende a **nueve mil sesenta y seis euros con setenta céntimos** (9066,70 €).

ANEXOS

Desarrollo de un algoritmo de aprendizaje automático para predicción de evolución de pacientes hospitalizados por COVID-19

Documento III

Belén Ballester Viciano

Grado en Ingeniería Biomédica

Curso académico 2020 – 2021

1. ANEXO I

Descripción: Tablas resumen de la información contenida en la base de datos de HM Hospitales.

- **Tabla 1ª**

ING/INPAT COVID CON/WITH URG/EMERG	**Registros de pacientes ingresados: demográficos, datos del episodio del ingreso, datos del episodio de Urgencias previo, si ha habido, datos del paso por UCI, si ha habido**
PATIENT ID	Identificador único de paciente ingresado
EDAD/AGE	Edad del paciente a abril del 2020
SEXO/SEX	Sexo del paciente
DIAG ING/INPAT	Diagnostico COVID durante el ingreso
F_INGRESO/ADMISSION_D_ING/INPAT	Fecha de ingreso como paciente tipo ingresado
F_ENTRADA_UC/ICU_DATE_IN	Fecha de entrada en la UCI durante el ingreso
F_SALIDA_UCI/ICU_DATE_OUT	Fecha de salida de la UCI durante el ingreso
UCI_DIAS/ICU_DAYS	Días en la UCI
F_ALTA/DISCHARGE_DATE_ING	Fecha de alta como paciente tipo ingresado
MOTIVO_ALTA/DESTINY_DISCHARGE_ING	Motivo/destino del alta como paciente tipo ingresado
F_INGRESO/ADMISSION_DATE_URG/EMERG	Fecha de la visita a Urgencias que derivó en ingreso
HORA/TIME_ADMISION/ADMISSION_URG/EMERG	Hora de la visita a la Urgencia que derivó en ingreso
ESPECIALIDAD/DEPARTMENT_URG/EMERG	Especialidad de la urgencia que derivó en ingreso

DIAG_URG/EMERG	Diagnóstico en el momento de la admisión en Urgencias
DESTINO/DESTINY_URG/EMERG	Destino de la urgencia (todas estas acabaron en ingresos porque el análisis es sobre pacientes ingresados)
HORA/TIME_CONSTANT_PRIMERA/FIRST_URG/EMERG	Hora del primer registro de constantes en la Urgencia
TEMP_PRIMERA/FIRST_URG/EMERG	Primer registro de temperatura tomado en la Urgencia
FC/HR_PRIMERA/FIRST_URG/EMERG	Primer registro de frecuencia cardiaca tomado en la Urgencia
GLU_PRIMERA/FIRST_URG/EMERG	Primer registro de glucemia tomado en la Urgencia
SAT_02_PRIMERA/FIRST_URG/EMERG	Primer registro de saturación de oxígeno tomado en la Urgencia
TA_MAX_PRIMERA/FIRST/EMERG_URG	Primer registro de tensión arterial mínima tomado en la Urgencia
TA_MIN_PRIMERA/FIRST_URG/EMERG	Primer registro de tensión arterial máxima tomado en la Urgencia
HORA/TIME_CONSTANT_ULTIMA/LAST_URG/EMERG	Hora del último registro de constantes en la Urgencia
FC/HR_ULTIMA/LAST_URG/EMERG	Ultimo registro de temperatura tomado en la Urgencia
TEMP_ULTIMA/LAST_URG/EMERG	Ultimo registro de frecuencia cardiaca tomado en la Urgencia
GLU_ULTIMA/LAST_URG/EMERG	Ultimo registro de glucemia tomado en la Urgencia
SAT_02_ULTIMA/LAST_URG/EMERG	Ultimo registro de saturación de oxígeno tomado en la Urgencia
TA_MAX_ULTIMA/LAST_URGEMERG	Ultimo registro de tensión arterial mínima tomado en la Urgencia
TA_MIN_ULTIMA/LAST_URG/EMERG	Ultimo registro de tensión arterial máxima tomado en la Urgencia

- **Tabla 2ª**

FARMACOS/DRUGS ING/INPAT (NO UCI/ICU)	**Registros de la medicación pautada y administrada durante el ingreso (no se recoge la medicación de UCI)**
PATIENT ID	Identificador único de paciente ingresado
FARMACO/DRUG_NOMBRE_COMERCIAL/COMERCIAL_NAME	Nombre comercial del fármaco
DOSIS_MEDIA_DIARIA/DAILY_AVRG_DOSE	Dosis media diaria administrada
INICIO_TRAT/DRUG_START_DATE	Inicio del tratamiento
FIN_TRAT/DRUG_END_DATE	Fin del tratamiento
ATC5_NOMBRE/NAME	Descripción de la categoría ATC5 en la clasificación ATC
ID_ATC5	Identificador de la categoría ATC5 en la clasificación ATC
ATC7_NOMBRE/NAME	Descripción de la categoría ATC7 en la clasificación ATC
ID_ATC7	Identificador de la categoría ATC7 en la clasificación ATC

- **Tabla 3ª**

CONSTANTS ING/INPAT (NO UCI/ICU)	**Registros de las constantes durante el ingreso (no se recogen las de UCI)**
PATIENT ID	Identificador único de paciente ingresado
CONSTANTS_ING/INPAT_FECHA/DATE	Fecha de registro de la constante
CONSTANTS_ING/INPAT_HORA/TIME	Hora de registro de la constante
FC/HR_ING/INPAT	Valor de frecuencia cardiaca
GLU/GLY_ING/INPAT	Valor de glucemia
SAT_02_ING/INPAT	Valor de saturación de oxígeno
TA_MAX_ING/INPAT	Valor de tensión arterial máxima
TA_MIN_ING/INPAT	Valor de tensión arterial mínima
TEMP_ING/INPAT	Valor de temperatura

- **Tabla 4ª**

LAB ING/INPAT - URG/EMERG	**Registros de los resultados de peticiones de laboratorio realizadas durante el ingreso y en la urgencia previa, si ha habido**
PATIENT ID	Identificador único de paciente ingresado
PETICION_LABORATORIO/LAB_NUMBER	Identificador de la petición de laboratorio (empiezan por "I" las del episodio de ingreso y por "U" las del episodio de Urgencias)
FECHA_PETICION/LAB_DATE	Fecha de la petición al laboratorio
HORA_PETICION/TIME_LAB	Hora de la petición al laboratorio
DETERMINACION/ITEM_LAB	Determinación
RESULTADO/VAL_RESULT	Resultado de la determinación
UNIDADES/UD_RESULT	Unidades del resultado de la determinación
VALORES_REFERENCIA/REF_VALUES	Valores de referencia para la determinación

- **Tabla 5ª**

CODIF CIE10/IDC URG/EMERG	**Registros de la codificación de la urgencia, si ha habido, según la CIE10**
PATIENT ID	Identificador único de paciente ingresado
DIA_PPAL	Diagnóstico Principal
DIA_02	Diagnóstico numero 2
DIA_03	Diagnóstico numero 3
DIA_04	Diagnóstico numero 4
DIA_05	Diagnóstico numero 5
DIA_06	Diagnóstico numero 6
DIA_07	Diagnóstico numero 7
DIA_08	Diagnóstico numero 8
DIA_09	Diagnóstico numero 9
DIA_10	Diagnóstico numero 10
DIA_11	Diagnóstico numero 11
DIA_12	Diagnóstico numero 12
PROC_01	Procedimiento número 1
PROC_02	Procedimiento número 2
PROC_03	Procedimiento número 3
PROC_04	Procedimiento número 4
PROC_05	Procedimiento número 5

- **Tabla 6ª**

CODIF CIE10/IDC10 ING/INPAT	**Registros de la codificación del ingreso según la CIE10**
PATIENT ID	Identificador único de paciente ingresado
DIA_PPAL	Diagnóstico Principal
DIA_02	Diagnóstico numero 2
DIA_03	Diagnóstico numero 3
DIA_04	Diagnóstico numero 4
DIA_05	Diagnóstico numero 5
DIA_06	Diagnóstico numero 6
DIA_08	Diagnóstico numero 7
DIA_07	Diagnóstico numero 8
DIA_09	Diagnóstico numero 9
DIA_10	Diagnóstico numero 10
DIA_11	Diagnóstico numero 11
DIA_12	Diagnóstico numero 12
DIA_13	Diagnóstico numero 13
DIA_14	Diagnóstico numero 14
DIA_15	Diagnóstico numero 15
DIA_16	Diagnóstico numero 16
DIA_17	Diagnóstico numero 17

DIA_18	Diagnóstico numero 18
DIA_19	Diagnóstico numero 19
NEO_01	Neoplasia número 1
NEO_02	Neoplasia número 2
NEO_03	Neoplasia número 3
NEO_04	Neoplasia número 4
NEO_05	Neoplasia número 5
NEO_06	Neoplasia número 6
POAD_02	POA (Present on Admission) para diagnóstico 2
POAD_03	POA (Present on Admission) para diagnóstico 3
POAD_04	POA (Present on Admission) para diagnóstico 4
POAD_05	POA (Present on Admission) para diagnóstico 5
POAD_06	POA (Present on Admission) para diagnóstico 6
POAD_07	POA (Present on Admission) para diagnóstico 7
POAD_08	POA (Present on Admission) para diagnóstico 8
POAD_09	POA (Present on Admission) para diagnóstico 9
POAD_10	POA (Present on Admission) para diagnóstico 10
POAD_11	POA (Present on Admission) para diagnóstico 11
POAD_12	POA (Present on Admission) para diagnóstico 12
POAD_13	POA (Present on Admission) para diagnóstico 13

POAD_14	POA (Present on Admission) para diagnóstico 14
POAD_15	POA (Present on Admission) para diagnóstico 15
POAD_16	POA (Present on Admission) para diagnóstico 16
POAD_17	POA (Present on Admission) para diagnóstico 17
POAD_18	POA (Present on Admission) para diagnóstico 18