UNIVERSIDAD DE VALLADOLID

FACULTAD DE MEDICINA

ESCUELA DE INGENIERÍAS INDUSTRIALES

TRABAJO DE FIN DE GRADO

GRADO EN INGENIERÍA BIOMÉDICA

Diseño y validación de un modelo de predicción de reingreso hospitalario por exacerbación de enfermedad pulmonar obstructiva crónica mediante métodos de aprendizaje computacional

Autora:

D.ª MARÍA TAMAYO POLO

Tutores:

D. DANIEL ÁLVAREZ GONZÁLEZ

D. TOMÁS RUIZ ALBI

Valladolid, septiembre de 2025

TÍTULO:	**Diseño y validación de un modelo de predicción de reingreso hospitalario por exacerbación de enfermedad pulmonar obstructiva crónica mediante métodos de aprendizaje computacional**
AUTORA:	**D.ª María Tamayo Polo**
TUTORES:	**D. Daniel Álvarez González** **D. Tomás Ruiz Albi**
DEPARTAMENTOS:	**Departamento de Teoría de la Señal y Comunicaciones e Ingeniería Telemática** **Departamento de Medicina, Dermatología y Toxicología**

TRIBUNAL

PRESIDENTE:	**D. Gonzalo C. Gutiérrez Tobal**
SECRETARIO:	**D. Daniel Álvarez González**
VOCAL:	**D. Tomás Ruiz Albi**
SUPLENTE 1:	**D. Mario Martínez Zarzuela**
SUPLENTE 2:	**D. Carlos Gómez Peña**

FECHA:	**septiembre de 2025**
CALIFICACIÓN:	

A Visi y Jesús,

refugios y guías de mi vida.

Agradecimientos

Este Trabajo de Fin de Grado supone el final de un camino emprendido hace cuatro años, tan maravilloso como desafiante, repleto de retos y momentos memorables. Me es imposible atribuirme el resultado de esta travesía a mi persona únicamente, porque nada de esto hubiera sido posible sin todos aquellos que han creído en mí cuando ni siquiera yo lo hacía y que me han tendido la mano en todos mis momentos de mayor tempestad. Por ello, quiero aprovechar este apartado para dedicarles unas palabras.

Me gustaría transmitir, en primer lugar, mi más sincero agradecimiento a mis tutores. A Daniel Álvarez González por su compromiso desde el minuto cero, su innegable preocupación porque todo saliera bien, su incansable dedicación, paciencia y consejos que me guiaron cuando perdía el rumbo a lo largo de este proyecto. Pero, sobre todo, por hacer lo difícil algo más fácil, transformando retos en pequeños pasos de la mano de su apoyo incondicional. A Tomás Ruiz Albi, porque no pude tener un recibimiento y trato más cálido durante mi estancia en el Servicio de Neumología. Por su inmenso afán porque cada día aprendiese algo nuevo en el hospital, su sabiduría, cercanía, entusiasmo por compartir todos sus conocimientos, implicación, generosidad y por hacerme sentir como en casa. Me siento profundamente agradecida por haber contado con estos tutores, un reflejo claro de que el verdadero valor de un profesional se basa en su humanidad.

Asimismo, quisiera dar las gracias a Fernando Moreno por su compañía, amabilidad e incondicional disposición para ayudarme siempre. Tampoco puedo olvidarme del personal médico y enfermeros del Servicio de Neumología, que me acogieron con los brazos abiertos e hicieron mi estancia más especial.

Deseo también expresar mi gratitud a aquellos compañeros que me han acompañado durante esta etapa, especialmente a María del Pilar y a Daniel. Me siento muy afortunada de haber podido contar con personas así y para mí son un regalo que me ha dado la carrera. Sin ellos, todo hubiera sido diferente y no recordaría este viaje con la misma sonrisa. Así que gracias de todo corazón.

A mi familia, pero sobre todo a mis padres, por ser el faro en mis días de tormenta, mi brújula cuando me sentía perdida y mi último aliento cuando no me quedaban fuerzas. Gracias por su amor incondicional, inmensa paciencia y por enseñarme el verdadero valor de la perseverancia y la constancia. Mi versión de hoy es el reflejo de las sabias lecciones y valores que me han inculcado y de las palabras que me alentaban a soñar y no rendirme nunca. Sin ellos, esto no tendría sentido, por lo que este trabajo es tanto mío como suyo.

Mención especial también para Luis, por ser mi mayor pilar durante todos estos años. Por ayudarme a remar contracorriente, ser un refugio cuando más lo necesitaba y, en definitiva, mi hogar.

Por último, y no por ello menos importante, quiero expresar mi agradecimiento a mis abuelos, por llenarme de cariño y dejar en mí una huella permanente. Me gustaría dedicar una mención especial a aquellos que ya no están, a mi abuelo José Luis, Boni y, en particular a mi abuela Puri, cuya bondad y ternura serán siempre recordadas y a la que le hubiera gustado saber hasta dónde he llegado. Este proyecto es también un tributo a ellos.

Gracias con todo mi corazón.

Resumen

Antecedentes. La Enfermedad Pulmonar Obstructiva Crónica (EPOC) es una patología respiratoria caracterizada por la limitación crónica al flujo aéreo, relacionada comúnmente con el hábito tabáquico. Afecta principalmente a la población mayor de 40 años y su prevalencia mundial en 2024 fue de 391 millones de personas. Según datos de la Organización Mundial de la Salud, actualmente constituye la cuarta causa de muerte en el mundo, pero se proyecta un aumento de su mortalidad para el año 2030, en el que podría ocupar el tercer puesto. Pese a la falta de consenso, las exacerbaciones pueden definirse como un aumento agudo de síntomas como disnea, tos o esputo. Sus consecuencias incluyen un empeoramiento de la función pulmonar, deterioro de la calidad de vida de los pacientes y aumento de mortalidad. Asimismo, representan una elevada carga económica y social, pero su infradiagnóstico persiste y no reciben la atención necesaria a pesar de su gran impacto. Esto subraya la necesidad de intervenciones urgentes con las que mitigar sus efectos, entre las que se encuentra la implementación de herramientas predictivas que permitan adoptar estrategias preventivas, efectuar un correcto manejo de las agudizaciones, abaratar costes y mejorar la vida de los pacientes y sus familiares.

Hipótesis y objetivos. El presente trabajo se elaboró bajo la hipótesis principal de que el desarrollo de modelos predictivos basados en técnicas de *Machine Learning* podría ser útil para estimar futuros reingresos por exacerbación de EPOC en los 30 días posteriores al alta. Con este fin, se esperaría que las variables recopiladas en el ámbito hospitalario permitiesen reconocer patrones asociados al reingreso. Como objetivo principal, se estableció desarrollar y validar un modelo predictivo de reingreso por exacerbación en un periodo de 30 días post-alta mediante técnicas de aprendizaje automático. Los objetivos específicos fueron: (i) recopilar una base de datos retrospectiva y prospectiva procedente del ámbito clínico, destinando la retrospectiva a la construcción y validación interna del modelo y la prospectiva a la validación temporal independiente del mismo; (ii) determinar las variables más relevantes para la predicción de reingresos por agudizaciones de EPOC; (iii) comparar el rendimiento predictivo de sendos enfoques de aprendizaje computacional: (a) *Random Forest*, propuesto como enfoque novedoso en el ámbito de la predicción de reingreso; y (b) red neuronal perceptrón multicapa, tomado como *benchmark* de referencia.

Materiales y métodos. En este estudio observacional ambispectivo de diseño y validación de modelos predictivos, se reclutaron pacientes procedentes del Servicio de Neumología del Hospital Universitario Río Hortega de Valladolid. El estudio constó de dos fases: una etapa retrospectiva, cuyos datos fueron adquiridos entre octubre de 2017 y junio de 2019, y otra prospectiva, que abarcó un periodo temporal de enero a junio de 2025. Los pacientes incluidos presentaban diagnóstico previo confirmado de EPOC e ingreso hospitalario por exacerbación de la patología. La metodología aplicada en este trabajo se dividió en las siguientes etapas: análisis descriptivo y curación del *dataset*, selección de características predictoras, diseño y optimización de modelos predictivos y validación temporal en una población prospectiva. En primera

instancia, se efectuó una imputación de datos aplicando *K* vecinos más cercanos. Tras esto, se seleccionó un subconjunto de las variables más relevantes y complementarias mediante el algoritmo *ReliefF*. A continuación, se realizó el diseño y optimización de un modelo predictivo bajo un enfoque de clasificación binaria: clase positiva (reingreso) o negativa (no reingreso). Para ello, se compararon sendos enfoques de aprendizaje computacional: *ensemble learning*, mediante *Random Forest* (RF) y redes neuronales, mediante un perceptrón multicapa (*Multi-Layer Perceptron*, MLP), cuyo rendimiento diagnóstico fue finalmente validado en una población prospectiva.

Resultados. 246 pacientes de la etapa retrospectiva cumplieron los criterios de inclusión, de los que finalmente se incluyeron 243 sujetos (42 reingresos y 201 no reingresos). En la base prospectiva, por su parte, 75 pacientes cumplieron los criterios de inclusión, incorporando finalmente 10 individuos (1 reingreso y 9 no reingresos). El algoritmo *ReliefF* seleccionó un total de 24 variables predictoras, destacando la pauta de mucolíticos al ingreso y alta, la presencia de microorganismos resistentes, el test TAI de uso correcto de inhaladores y la prescripción de oxigenoterapia domiciliaria basal. En la base de datos retrospectiva, *Random Forest* presentó una sensibilidad del 50.0%, especificidad del 91.7%, precisión o *accuracy* del 84.7%, *F1 score* del 52.2% y AUC de 0.826 en un subconjunto de pacientes de *test*. La red neuronal obtuvo una sensibilidad del 75.0%, especificidad del 85.0%, precisión o *accuracy* del 83.3%, *F1 score* del 60.0% y AUC de 0.857. En la cohorte prospectiva, *Random Forest* logró una sensibilidad del 100%, especificidad del 66.7%, precisión o *accuracy* del 70.0% y *F1 score* del 40.0%. La red neuronal alcanzó una sensibilidad del 100%, especificidad del 55.6%, precisión o *accuracy* del 60.0% y *F1 score* del 33.3%. MLP resultó ser superior en la base de datos retrospectiva a *Random Forest* en términos de *F1 score* y AUC, aunque su especificidad alcanzó un valor destacable. En la base de datos prospectiva, el modelo basado en *Random Forest* mostró mayor capacidad de generalización que la red neuronal MLP.

Conclusiones. Los modelos predictivos desarrollados basados en *Random Forest* y MLP mostraron una capacidad predictiva notable para la estimación de reingresos por exacerbación de EPOC en los 30 días posteriores al alta. Ambos modelos alcanzaron valores de AUC superiores a 0.8, aunque la red MLP mostró un rendimiento mayor. El algoritmo *ReliefF* seleccionó un conjunto de variables que demostraron su utilidad en la identificación de patrones relacionados con el reingreso, destacando a su vez la heterogeneidad de la enfermedad. La validación temporal confirmó la viabilidad de los modelos en nuevas cohortes, aunque la fiabilidad de sus métricas aumentaría con bases de datos prospectivas más amplias. En comparación con otras investigaciones previas, los rendimientos predictivos alcanzaron resultados competitivos, incluso con un tamaño muestral limitado.

Palabras clave

Enfermedad Pulmonar Obstructiva Crónica, reingreso hospitalario, exacerbación, *Machine Learning*, *ReliefF*, *Random Forest*, red neuronal perceptrón multicapa.

Abstract

Background. Chronic obstructive pulmonary disease (COPD) is a respiratory condition characterised by progressive airflow limitation, commonly associated with smoking. It mainly affects people over the age of 40, and its global prevalence in 2024 was 391 million people. According to the World Health Organization data, it is currently the fourth leading cause of death worldwide, but mortality is projected to increase by 2030, when it could rank third. Despite the lack of consensus, exacerbations can be defined as an acute increase in symptoms such as dyspnoea, coughing or sputum. Their consequences include worsening lung function, deterioration in patients' quality of life and increased mortality. They also represent a high economic and social burden, but they remain underdiagnosed and do not receive the necessary attention despite their significant impact. This highlights the need for urgent interventions to mitigate their effects, including the implementation of predictive tools that enable preventive strategies to be adopted, exacerbations to be managed correctly, costs to be reduced, and the lives of both patients and their families to be improved.

Hypothesis and objectives. This study was conducted under the main hypothesis that the development of predictive models based on machine learning techniques could be useful to estimate future readmissions due to COPD exacerbation within 30 days of discharge. To this end, it was expected that the variables collected in the hospital setting would allow patterns associated with readmission to be recognised. The main objective was to develop and validate a predictive model of readmission due to exacerbation within 30 days of discharge using machine learning techniques. The specific objectives were: (i) to build both retrospective and prospective databases from the clinical setting, with the retrospective data being used for the construction and internal validation of the model and the prospective data for its independent temporal validation; (ii) to determine the most relevant variables for predicting readmissions due to COPD exacerbations; (iii) to compare the predictive performance of two machine learning approaches: (a) Random Forest, proposed as a novel approach in the field of readmission prediction; and (b) multilayer perceptron neural network, taken as a benchmark.

Materials and methods. In this observational, ambispective study of predictive model design and validation, patients were recruited from the Pulmonology Department of the Río Hortega University Hospital in Valladolid. The study consisted of two phases: a retrospective stage, whose data were acquired between October 2017 and June 2019, and a prospective stage, which covered a period from January to June 2025. The patients included had a previous confirmed diagnosis of COPD and were admitted to hospital due to exacerbation of the disease. The methodology applied in this study was divided in the following stages: descriptive analysis and data curation of the dataset, selection of predictive variables, design and optimisation of predictive models, and temporal validation in a prospective population. First, data imputation was performed using K nearest neighbours. After this, a subset of the most relevant and complementary variables was selected using the ReliefF algorithm. Next, a predictive model was

designed and optimised using a binary classification approach: positive class (readmission) or negative class (no readmission). To this end, two machine learning approaches were compared: ensemble learning, using Random Forest (RF), and neural networks, using a multi-layer perceptron (MLP), whose predictive performance was finally validated in a prospective population.

Results. A total of 246 patients from the retrospective stage met the inclusion criteria, of whom 243 subjects were finally included (42 readmissions and 201 non-readmissions). In the prospective database, 75 patients met the inclusion criteria, with 10 individuals finally being included (1 readmission and 9 non-readmissions). The ReliefF algorithm selected a total of 24 predictor variables, highlighting the pattern of mucolytics at admission and discharge, the presence of resistant microorganisms, the TAI test for correct use of inhalers, and the prescription of baseline home oxygen therapy. In the retrospective database, Random Forest had a sensitivity of 50.0%, specificity of 91.7%, accuracy of 84.7%, F1 score of 52.2%, and AUC of 0.826. The neural network obtained a sensitivity of 75.0%, specificity of 85.0%, accuracy of 83.3%, F1 score of 60.0%, and AUC of 0.857. In the prospective cohort, Random Forest achieved a sensitivity of 100%, specificity of 66.7%, accuracy of 70.0% and F1 score of 40.0%. The neural network achieved a sensitivity of 100%, specificity of 55.6%, accuracy of 60.0% and F1 score of 33.3%. MLP proved to be superior to Random Forest in the retrospective database, although its specificity was noteworthy. In the prospective database, the Random Forest model showed higher generalisation ability compared to MLP.

Conclusions. The predictive models developed based on Random Forest and MLP showed remarkable predictive capacity for estimating readmissions due to COPD exacerbation within 30 days after discharge. Both models have AUCs greater than 0.8, although the MLP network reached higher performance. The ReliefF algorithm selected a set of variables that proved to be useful in identifying patterns related to readmission, while also highlighting the heterogeneity of the disease. Temporal validation confirmed the viability of the models in new cohorts, although the reliability of their metrics would increase with larger prospective databases. Compared to other previous research, the predictive performance achieved competitive results, even with a limited sample size.

Keywords

Chronic Obstructive Pulmonary Disease, hospital readmission, exacerbation, Machine Learning, ReliefF, Random Forest, multilayer perceptron neural network.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Glosario de siglas y acrónimos

A Muco	Mucolíticos al alta
A Teo	Teofilinas al alta
A_Min_SABA	Medicación inhaladora al alta, SABA
A_Mre_SABA	Medicación de rescate al alta, SABA
Acc	Precisión o *accuracy*
ACV	Accidente cerebrovascular
ADN	Ácido desoxirribonucleico
AECOPD	Exacerbaciones agudas de EPOC
alpha	Parámetro de regularización del MLP
AMP	*Amplitude*
AOS	Apnea obstructiva del sueño
AUC	Área bajo la curva ROC
AVAD (DALY en inglés)	Años de Vida Ajustados por Discapacidad
AVD	Años vividos con discapacidad
AVP	Años de vida perdidos
B Muco	Mucolíticos en estado basal
B Oxi Dom	Oxigenoterapia continua domiciliaria basal
B Teo	Teofilinas en estado basal
B_Esp_FEV1	FEV1 basal
B_Esp_FEV1_p	FEV1 basal (% teórico)
B_Esp_FVC_p	FVC basal (% teórico)
B_Mre_CI	Medicación de rescate basal, corticoides
B_Mre_SABA	Mediación de rescate basal, SABA
$\boldsymbol{b_j}$	Salida de la neurona oculta "j" en el MLP.
BP	*Blood pressure*
BPC	Buenas Prácticas Clínicas
BT	*Blood test*
CAT	*COPD Assesment Test*
CEIm	Comité de Ética de la Investigación con medicamentos
CNAF	Cánula nasal de alto flujo
Comp Arritmias	Complicación de arritmias
Comp Neumonía	Complicación de neumonía
COVID-19	Virus SARS-CoV-2
CPAP	Presión positiva continua de aire
CVRS	Calidad de vida relacionada con la salud
DAAT	Déficit de α-1 antitripsina
DL	*Deep Learning*
Edemas Per	Edemas periféricos
EGFR	Factor de crecimiento epidérmico
EMR	*Electronic health records*
EPOC (COPD en inglés)	Enfermedad Obstructiva Pulmonar Crónica
EuroQOL-5D	*European Quality of Life 5 Dimensions*
F	Función de activación en el MLP.
FEV1	Volumen espiratorio forzado en el primer segundo
FiO_2	Fracción inspirada de oxígeno

FN	Falsos negativos
FP	Falsos positivos
FPR	Tasa de falsos positivos
FREQ	*Frequency*
FVC	Capacidad vital forzada
GBD	*Global Burden of Disease*
Gen AI	*Generative AI*
GesEPOC	Guía Española de la EPOC
GOLD	*Global Initiative for Chronic Obstructive Lung Disease*
GRC	Grupo de riesgo clínico
H_0	Hipótesis nula
H_1	Hipótesis alternativa
HCO_3	Bicarbonato
HR	*Heart rate*
I Muco	Mucolíticos en el ingreso
I Teo	Teofilinas en el ingreso
I VNI	Ventilación no invasiva al ingreso
I_Gas_HCO3	HCO_3 al ingreso
I_Gas_PCO2	PCO_2 al ingreso
I_Min_LABA	Medicación inhaladora al ingreso, LABA
I_Min_LAMA	Medicación inhaladora al ingreso, LAMA
I_Min_SABA	Medicación inhaladora al ingreso, SABA
IA	Inteligencia Artificial
IAM	Infarto agudo de miocardio
ICS	*Corticoides inhalados*
IFDE4	Inhibidores de fosfodiesterasa tipo 4
ILC2	Células linfoides innatas tipo 2
IMC	Índice de Masa Corporal
INE	Instituto Nacional de Estadística
IQR	Rango intercuartílico
k_j	Salida de la neurona "j" aplicando la función de activación en el MLP.
KNN	*K-Nearest Neighbors*
LABA	*Long Acting Beta Agonists*
LAMA	*Long Acting Muscarinic Antragonists*
LR-	Razón de verosimilitud negativa
LR+	Razón de verosimilitud positiva
LS	*Lung sounds*
MaxNumSplits	Máximo número de divisiones en *Random Forest*
miARN	Micro-ácido oxirribonucleico
Microorg Resis	Microorganismos resistentes
MinLeafSize	Tamaño mínimo de hoja en *Random Forest*
ML	*Machine Learning*
MLP	*Multi-Layer Perceptron*
mMRC	*modified Medical Research Council*
Mov Torácicos	Movimientos torácicos paradójicos
MT	Máquina de Turing
MUC5AC	Glicoproteína secretora de mucina-5AC
N Ingresos	Número de ingresos por agudización (año previo)

N_Días_Ing	Número de días ingresado
N_Ingresos	Número de ingresos por agudización (año previo)
NaN	*Not a Number*
net_j	Entrada neta a la neurona "j" de la capa oculta.
net_k	Entrada neta a la neurona "k" de la capa de salida.
NPV	Valor predictivo negativo
NumPredictorsToSample	Número de predictores a muestrear en *Random Forest*
NumTrees	Número de árboles en *Random Forest*
OBB	*Out of the bag*
OMS	Organización Mundial de la Salud
$PaCO_2$	Presión de dióxido de carbono en sangre arterial
PaO_2	Presión de oxígeno arterial
PCO_2	Presión parcial de dióxido de carbono en la sangre
PCR	Proteína C reactiva
PIBM	Países con ingresos bajos y medianos
PM	Materia particulada
PO_2	Presión parcial de oxígeno en la sangre
PPV	Valor predictivo positivo
Q1	Primer cuartil
Q2	Segundo cuartil
Q3	Tercer cuartil
RF	*Random Forest*
ROC	*Receiver Operating Characteristic*
RR	*Respiratory rate*
RSV	Virus respiratorio sincitial
SABA	*Short Acting Beta Agonists*
SAMA	*Short Acting Muscarinic Antagonists*
SaO_2	Saturación de oxígeno en sangre arterial
SDRA	Síndrome de distrés respiratorio agudo
Se	Sensibilidad
SEPAR	Sociedad Española de Neumología y Cirugía Torácica
s/n	Sin número
Sp	Especificidad
SVM	*Support Vector Machine*
TAC	Tomografía axial computarizada
TAI	Test de Adhesión a los Inhaladores
TC	Tomografía computarizada
TEP	Tromboembolismo pulmonar
Test_E5D_Aco	Test EuroQoL-5D, actividades cotidianas
Test_E5D_CPe	Test EuroQoL-5D, cuidado personal
Test_E5D_Dep	Test EuroQoL-5D, depresión
Test_E5D_Dol	Test EuroQoL-5D, dolor
Test_TAI_Ade	Test de TAI, nivel de adhesión a los inhaladores
Test_TAI_I_Del	Test de TAI, incumplimiento deliberado de la pauta del inhalador
Test_TAI_I_Err	Test de TAI, incumplimiento errático de la pauta del inhalador
Test_TAI_I_Inc	Test de TAI, incumplimiento inconsciente de la pauta del inhalador
TN	Verdaderos negativos
TP	Verdaderos positivos

TPR	Tasa de verdaderos positivos
TVP	Trombosis venosa profunda
UE	Unión Europea
USD	Dólares estadounidenses
Uso Muscul Acce	Uso de musculatura accesoria
UVI	Unidad de Ventilación Invasiva
VAS	Escala análoga visual
VEGF	Factor de crecimiento del endotelio vascular
v_{kj}	Peso de la conexión entre la neurona "j" de la capa oculta y neurona "k" de la capa de salida en el MLP.
VNI	Ventilación mecánica no invasiva
W[A]	Peso calculado para la variable *A* por *ReliefF*
w_{ji}	Peso de la conexión entre neurona "i" de la capa de entrada y neurona "j" de la capa oculta en el MLP.
XAI	*Explainable Artificial Intelligence*
XGBoost	*Extreme Gradient Boosting*
x_i	Entrada a la neurona "i" de la capa de entrada al MLP.
y_k	Salida de la neurona "k" de la capa de salida en el MLP.
6MWT	*Six Minutes Walking Test*
α	Nivel de significancia estadística
ϑ_j	Sesgo para el ajuste de la activación de la neurona en el MLP.

CAPÍTULO 1. INTRODUCCIÓN

La Organización Mundial de la Salud (OMS) define la EPOC (*Enfermedad Pulmonar Obstructiva Crónica)* como una patología pulmonar caracterizada por la disminución persistente del flujo aéreo ocasionando múltiples problemas respiratorios [1]. Suele manifestarse en personas mayores de 40 años con antecedentes de tabaquismo [2] y cursa con síntomas como tos con esputo, disnea, sibilancias y fatiga [1]. A esta sintomatología hay que añadirle un mayor riesgo a desarrollar otras afecciones, entre las que se encuentran arritmias, insuficiencia cardíaca, cardiopatía isquémica, derrame pleural, neumonía, tromboembolismo pulmonar (TEP), neumotórax, síndrome de distrés respiratorio agudo (SDRA), sepsis, cáncer de pulmón, debilidad muscular, osteoporosis, depresión, ansiedad y atrofia muscular [1], [3]. Además de su gran carga clínica, cabe destacar el significativo impacto socioeconómico de los reingresos por exacerbación de la enfermedad y la repercusión sobre la calidad de vida de los pacientes y sus familiares [4]. En 2024, es estimó que el 75% de los casos de EPOC no eran identificados en España, constituyendo una de las patologías más infradiagnosticadas [5].

Todo lo expuesto subraya la necesidad de disponer de herramientas que permitan predecir el riesgo de reingreso hospitalario por agudización de EPOC, ya que estas contribuirían a adoptar medidas preventivas, optimizar recursos, mejorar el pronóstico del paciente y potenciar tanto su bienestar como el de su entorno. Por ello, el presente trabajo se basa en la elaboración de modelos predictivos destinados a estimar la probabilidad de reingreso por exacerbación, tratando así de proporcionar un aporte complementario que pueda enriquecer este ámbito.

1.1. Estructura del documento

La estructura del presente TFG se detalla a continuación, garantizando una mayor compresión global del mismo mediante una presentación ordenada y coherente de su contenido.

- **Capítulo 1: Introducción.** Se explican diversos aspectos de la EPOC, como sus características clínicas y pronóstico (fisiopatología, formas clínicas y mortalidad), etiología, factores de riesgo, diagnóstico, severidad y tratamiento. Tras un conocimiento más amplio de la patología, se expone el concepto de “exacerbación de EPOC”, abordando sus implicaciones, diagnóstico, gravedad, reingresos, mortalidad, factores predisponentes, medidas preventivas y volumen global. También se presenta el impacto socioeconómico de estas agudizaciones, resaltando la gran carga que suponen en gran diversidad de ámbitos y la necesidad urgente de implementar soluciones a un problema en la mayoría de los casos ignorado. Finalmente, se incluye una introducción a la Inteligencia Artificial (IA) y su presencia tanto en la vida cotidiana como en la EPOC.

- **Capítulo 2: Hipótesis y objetivos.** Se plantean las hipótesis bajo las que se sustenta este trabajo, tanto de carácter clínico como técnico. Estas pretenden ser corroboradas a través de los objetivos establecidos (principal y específicos).

- **Capítulo 3: Sujetos y variables de estudio.** Se exponen los aspectos éticos pertinentes para el desarrollo de este TFG, así como los criterios de inclusión y exclusión, tamaños muestrales de las bases de datos, información relativa a la adquisición de las cohortes en el ámbito hospitalario y aspectos relacionados con las variables de estudio.

- **Capítulo 4: Metodología.** Se describen los distintos métodos aplicados, profundizando en el tratamiento e imputación de datos faltantes, análisis descriptivo de las variables, selección de las características predictoras más relevantes y desarrollo de los modelos predictivos y su evaluación mediante métricas de rendimiento.

- **Capítulo 5: Resultados.** Se presentan los resultados obtenidos en cada fase descrita en el anterior capítulo.

- **Capítulo 6: Discusión.** Se analizan los hallazgos mostrados en los resultados, debatiendo su coherencia y tratando de comprender los motivos que han conducido a su obtención. Asimismo, se integra un enfoque autocrítico abordando las posibles limitaciones del estudio.

- **Capítulo 7: Conclusiones.** En este último capítulo, se exponen las principales contribuciones y conclusiones del trabajo, así como futuras líneas de investigación con las que contribuir a un mayor alcance del estudio.

1.2. Características clínicas y pronóstico de la EPOC

1.2.1. Fisiopatología

Los procesos fisiopatológicos de la EPOC consisten en alteraciones en las vías respiratorias, el parénquima pulmonar y su vasculatura. En este sentido, se experimenta inflamación, daños estructurales, limitación al flujo aéreo, disbiosis e infección [4], [3].

- **Inflamación**: se produce ante exposiciones inhalatorias, como el humo del tabaco. Estos irritantes provocan un incremento de proteasas y una reducción de antiproteasas. En condiciones fisiológicas normales, la función de las proteasas es la degradación de la elastina y el tejido conectivo para facilitar la reparación tisular. Las antiproteasas, como

la α-1 antitripsina, sirven como mecanismo de compensación para mantener el equilibrio [3].

Sin embargo, en la EPOC se observa un incremento del número de macrófagos, neutrófilos activados y linfocitos. La acción conjunta de estas células inflamatorias, junto con células epiteliales y estructurales, provoca la liberación de mediadores inflamatorios. Esto conlleva la atracción de células inflamatorias desde la circulación, con la consiguiente liberación excesiva de proteasas y, finalmente, daño estructural [3], [4].

Esta inflamación puede ser también sistémica, hecho que podría explicar la existencia de comorbilidades presentes en la EPOC. Cabe destacar que, en algunos pacientes, el patrón inflamatorio incluye un aumento de eosinófilos y células linfoides innatas tipo 2 (ILC2), semejante al de la enfermedad asmática [4].

El estrés oxidativo es otro aspecto importante a tener en cuenta. Los niveles elevados de neutrófilos y macrófagos, así como la inhalación de partículas nocivas, causan la acumulación de radicales libres, aniones superóxido y peróxido de hidrógeno. Este último, junto con la 8-isoprostano, es un biomarcador de estrés oxidativo y se halla elevado en el aliento exhalado, el esputo y la circulación sistémica. Estos oxidantes, sumados al descenso del nivel del factor de crecimiento del endotelio vascular (VEGF) y la liberación de neuropéptidos profibróticos, promueven la apoptosis del parénquima pulmonar [4], [3].

La inflamación puede resultar tan severa que, incluso en los perfiles clínicos más graves que han abandonado el consumo de tabaco, el proceso inflamatorio no se revierte por completo [3].

- **Alteraciones estructurales**: el desequilibrio previamente mencionado entre los niveles de proteasas derivadas de células inflamatorias y epiteliales, y las antiproteasas, favorece la degradación de la elastina del tejido conectivo. Además, tras la inflamación puede desencadenarse una fibrosis peribronquiolar y producirse múltiples lesiones en la pared de la vía aérea. Esto se traduce en un exceso de tejido muscular y fibroso, pudiendo incluso generarse obstrucciones respiratorias. Incluso en pacientes con EPOC leve se han apreciado cambios estructurales en la vasculatura pulmonar [4].

- **Limitación al flujo de aire**: la hipersecreción de moco debida a la inflamación, los tapones mucosos, el broncoespasmo y la fibrosis peribronquial son responsables de la disminución de la superficie útil de las vías aéreas y de su obstrucción. Además, el parénquima pulmonar pierde su adherencia como consecuencia de la destrucción de los tabiques alveolares. Todo ello conlleva un aumento del trabajo respiratorio como mecanismo de compensación ante el incremento de resistencia de las vías aéreas. El resultado es una hipoventilación alveolar con hipoxia e hipercapnia [3].

- **Disbiosis**: se refiere a la alteración del microbioma observada en pacientes con EPOC. La exposición a factores de riesgo provoca daños en el microbioma intestinal y respiratorio, además de afectar la inmunidad mucosa e inducir inflamación pulmonar a través de respuestas inmunitarias. El microbioma varía tras padecer una infección viral, durante las exacerbaciones y en respuesta al uso de antibióticos y corticosteroides, tanto orales como inhalados [4].

- **Infección**: los procesos infecciosos agravan la inflamación y, en consecuencia, afectan negativamente la evolución de la enfermedad. La infección se ve favorecida por la dificultad para eliminar el moco en las vías aéreas inferiores. *Haemophilus influenzae* se detecta en aproximadamente el 30% de los sujetos con EPOC, mientras que *Pseudomonas aeruginosa* y otras bacterias gramnegativas se encuentran en fases más avanzadas [3].

1.2.2. Formas clínicas

Aunque EPOC se emplea como término general, este engloba dos entidades: enfisema y bronquitis crónica. En gran parte de los pacientes que padecen la enfermedad, coexisten ambos trastornos, por lo que se acuña el concepto de EPOC para una mayor precisión [2].

El enfisema se presenta como una destrucción del parénquima pulmonar, en particular de las paredes alveolares, lo que provoca la pérdida del retroceso elástico pulmonar. En consecuencia, se incrementa el riesgo de colapso de las vías aéreas debido a la pérdida de tabiques alveolares y disminución de la tracción radial. Además, se produce una hiperinsuflación pulmonar y atrapamiento de aire. La morfología alveolar se ve alterada y adquieren flacidez. Todo ello conduce a un agrandamiento patológico de los espacios aéreos, una reducción del intercambio gaseoso y, en ocasiones, el desarrollo de bullas [2], [3].

En la Figura 1 puede apreciarse una comparativa esquemática entre un acino pulmonar sano y uno enfermo por enfisema [6]. En este sentido, se ilustran unos alvéolos debilitados y colapsados en contraste a la estructura normal en situación no patológica.

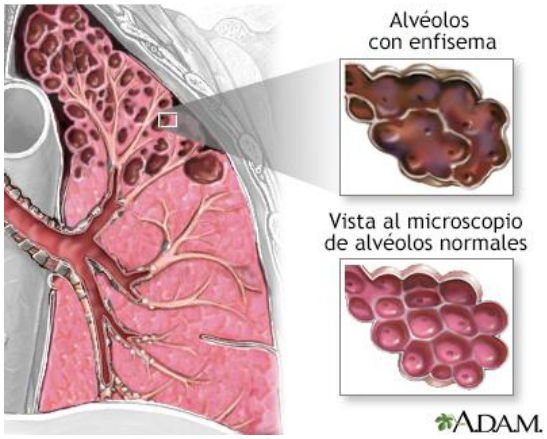

Figura 1. Comparativa entre sacos alveolares sanos (abajo) y patológicos por enfisema (arriba) [6].

Por su parte, la bronquitis crónica se manifiesta con irritación e inflamación de los conductos bronquiales. Esto se traduce en un engrosamiento del epitelio y producción excesiva de mucosidad espesa [2]. Se define por una tos productiva la mayoría de los días de la semana durante al menos tres meses al año, en dos años sucesivos [3]. En ocasiones, puede observarse la presencia de pus y fosas bronquiales prominentes en las aberturas de las glándulas mucosas bronquiales. Las células inflamatorias activan el receptor del factor de crecimiento epidérmico (EGFR), provocando la transcripción del gen de mucina (MUC5AC) que desencadena en una hipersecreción de moco [7] (véase Figura 2).

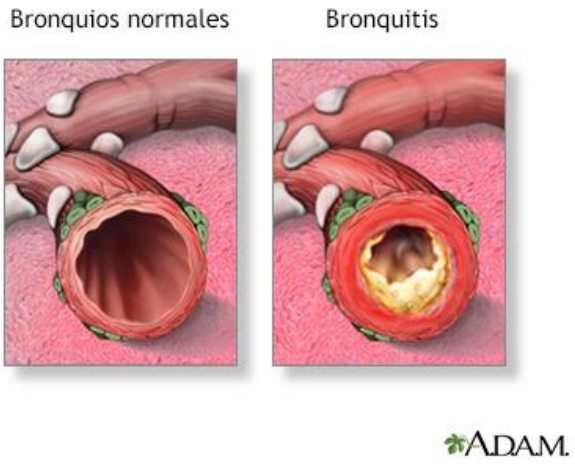

Figura 2. Comparativa entre bronquios sanos y bronquitis crónica [8].

En 1955, Dornhorst establece unos rasgos diferenciadores entre el enfisema y la bronquitis crónica. El paciente que padecía bronquitis crónica era llamado “Abotagado Azul” *(Blue Boaters)* y se caracteriza por cianosis, obnubilación, hematocrito ≥ 60%, edemas e insuficiencia cardíaca.

En cuanto al sujeto enfisematoso, se le denominaba como "Soplador Rosado" *(Pink Puffer)* con una notoria pérdida de peso, disnea progresiva y hematocrito < 55% [9]. En la Figura 3, puede observarse el aspecto físico típico de un paciente que sufre bronquitis crónica (izquierda) y enfisema (derecha).

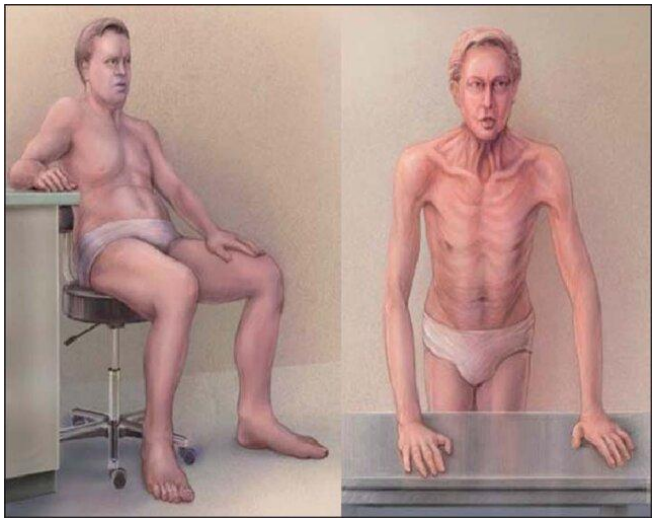

Figura 3. Rasgos físicos de paciente con bronquitis crónica vs. enfisema [9].

1.2.3. Mortalidad

Según datos de la OMS, la EPOC constituye la cuarta causa de muerte a nivel mundial. En 2021 fue responsable de aproximadamente 3.5 millones de defunciones, lo que equivale al 5% de fallecimientos globales [1]. Sin embargo, estas cifras deben interpretarse con precaución, puesto que la EPOC es una patología infradiagnosticada con frecuencia [10], lo que podría implicar una subestimación de la mortalidad real. Además, la OMS prevé que en el año 2030 la EPOC se convierta en la tercera causa de fallecimiento en el mundo, suponiendo el 7.8% de las defunciones totales y el 27% de las asociadas al tabaquismo. Solo sería superada por el cáncer (33%) y las enfermedades cardiovasculares (29%) [10].

El incremento de la mortalidad global se atribuye principalmente al aumento del tabaquismo en países de ingresos bajos y medianos, al envejecimiento de la población mundial y a la ausencia de tratamientos capaces de alterar el transcurso de la enfermedad [4]. De acuerdo con los datos actuales disponibles, se espera que en el año 2060 haya 5.4 millones de muertes anuales atribuibles a la EPOC [4].

En el caso de España, la tasa de mortalidad por EPOC por cada 100 000 habitantes muestra una tendencia distinta a la global, con un descenso sostenido tanto en hombres como en mujeres. Esto puede observarse en la Figura 4, donde se ilustra la mortalidad por sexos en España entre 1980 y 2023 [11]. No obstante, hay una diferencia muy significativa entre ambos sexos,

presentando los varones unas cifras de mortalidad notablemente superiores a las de las mujeres durante todo el intervalo temporal considerado. Esta disparidad podría explicarse por el consumo históricamente más elevado de tabaco entre los hombres.

La curva correspondiente a la población masculina muestra un descenso sostenido desde finales de los años 90. En contraposición, la evolución en las mujeres es más estable y presenta un descenso mucho más leve desde 1998 aproximadamente.

Dado que la EPOC está estrechamente vinculada al consumo de tabaco, los patrones representados en la Figura 4 pueden interpretarse como el reflejo de las variaciones en los hábitos tabáquicos de la población. En ambos sexos, el descenso de la mortalidad podría asociarse a una mayor concienciación sobre los efectos nocivos del tabaquismo. Sin embargo, la relativa estabilización en la población femenina, junto al aumento del número de mujeres fumadoras en los últimos años, alerta sobre la posibilidad de un repunte en el futuro.

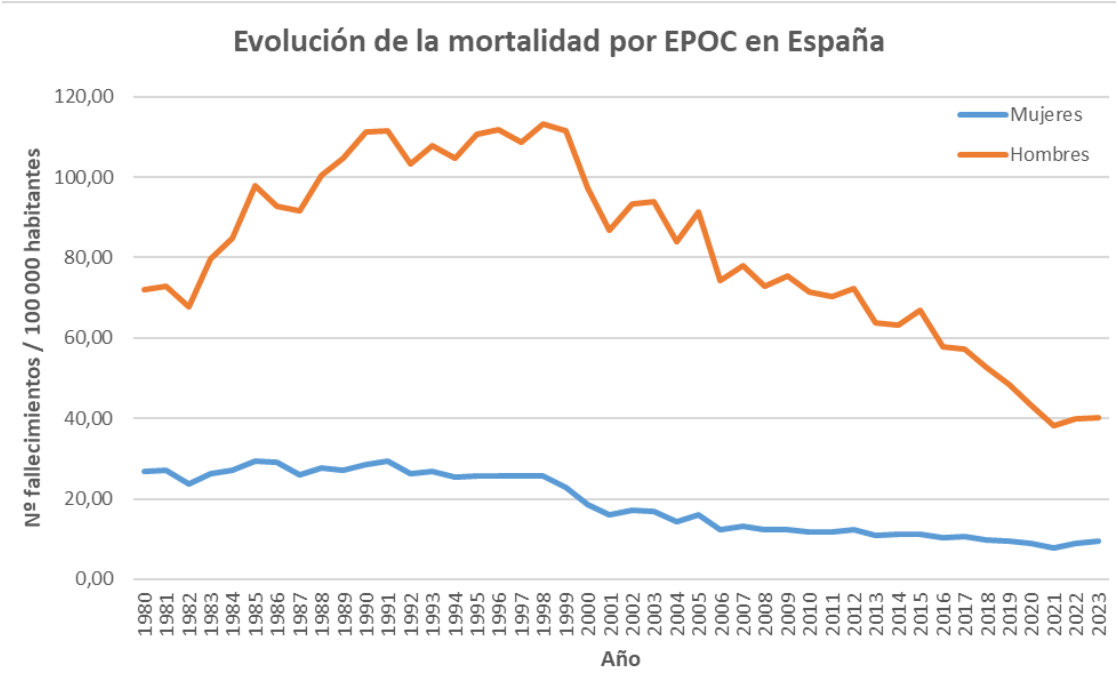

Figura 4. Tasa de mortalidad por EPOC en España de hombres y mujeres entre 1980 y 2023. Los datos fueron facilitados por el *Informe del Estado de Salud de la Población* (IESP) de la Comunidad de Madrid [11], los cuales provienen del Instituto Nacional de Estadística (INE).

1.3. Etiología y factores de riesgo

La EPOC es una enfermedad con un amplio abanico de factores de riesgo que contribuyen a su desarrollo. Estos factores pueden agruparse en dos categorías: ambientales y del huésped, y genéticos y epigenéticos [12]. Los principales son recogidos en la Figura 5 [13]:

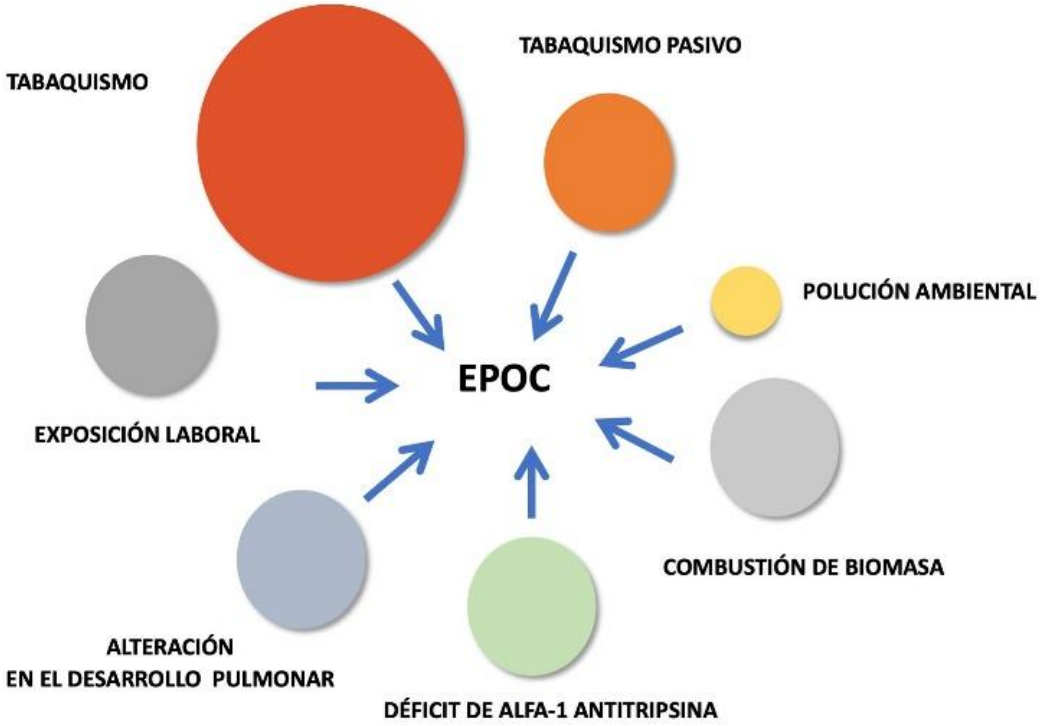

Figura 5. Principales factores de riesgo en EPOC [13].

1.3.1. Factores de riesgo ambientales y del huésped

- **Tabaquismo.**
  - **Activo.** Es el principal factor desencadenante de EPOC. El hábito tabáquico es responsable de una disminución del volumen espiratorio forzado en el primer segundo (FEV1) respecto a los pacientes no fumadores. Además, su alta nocividad puede causar bronquiolitis con aumento de producción mucosa, daño en paredes alveolares y fibrosis intersticial [12]. Estos daños estructurales y funcionales son clave en la aparición de la enfermedad.

  - **Pasivo.** Aunque los efectos y consecuencias del tabaquismo activo están ampliamente documentados, la exposición pasiva al humo del tabaco puede ocasionar la enfermedad [13]. Estudios como el de Tan et al. (2015) [14], esclarecen unos resultados donde se refleja cómo el tabaquismo pasivo afecta principalmente a mujeres. Sin embargo, en una población formada en su totalidad por pacientes fumadores, no existen diferencias entre sexos (Figura 6).

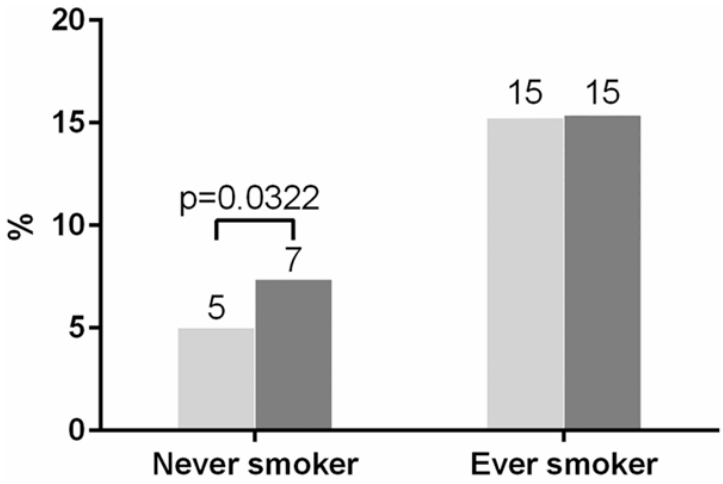

Figura 6: Comparativa entre sexos (color claro = hombre, color oscuro = mujer) en grupos de pacientes fumadores pasivos (never smoker) y activos (ever smoker) [14].

- **Polución ambiental.** La exposición a monóxido de carbono, óxidos de nitrógeno y sulfuro, así como material particulado (PM 2.5, PM 10), contribuye al desarrollo de alteraciones pulmonares como asma y EPOC [12].

- **Combustión de biomasa.** La quema de materia orgánica continúa siendo una fuente de energía en muchas viviendas, ya sea como método de calefacción o para actividades culinarias. Se estima que en torno al 50% de hogares a nivel mundial, y el 90% de las viviendas rurales, efectúan esta práctica [12]. Esto se traduce en un total de 3000 millones de personas expuestas globalmente al humo de esta combustión [12]. Además, el aumento sostenido de los costes de combustibles limpios, incluso en países desarrollados, incentiva su uso a pesar de los riesgos para la salud que supone.

  Se ha demostrado que la combustión de biomasa provoca una mayor limitación al flujo aéreo y debe evitarse especialmente en etapas tempranas de la vida con el fin de no condicionar el desarrollo pulmonar [13], [12].

- **Exposición laboral**. Los productos químicos, vapores y polvos orgánicos e inorgánicos presentes en actividades ocupacionales promueven la aparición de la enfermedad. Esto ha sido respaldado en diversos estudios, destacando el de Marchetti et al. (2014) [15], que constata una elevada limitación al flujo aéreo, desarrollo de enfisema y atrapamiento aéreo.

1.3.2. Factores de riesgo genéticos y epigenéticos

- **Déficit de α-1 antitripsina (DAAT).** Es una alteración congénita causante de enfisema en adultos. Aunque es el factor de riesgo genético más investigado, continúa siendo infradiagnosticado. El DAAT provoca una pérdida acelerada del parénquima pulmonar [13]. En cuanto a su prevalencia, esta condición es responsable de 1 de cada 700 casos de EPOC en el sur europeo y 1 de cada 4500 personas caucásicas se ven afectadas por la misma [16].

- **Modificaciones epigenéticas.** Son cambios químicos en el ADN que no dañan su secuencia, pero afectan directamente al genoma. Las exposiciones ambientales pueden inducir este fenómeno y en ocasiones son heredables. Entre los mecanismos más destacables pueden citarse [12]:
  - **Metilación del ADN:** adición de un grupo metilo a las citosinas adyacentes a guaninas.
  - **Alteración de histonas:** acetilación y metilación.
  - **Cambios en los micro-ARN (miARN).**

  Cuando los cambios epigenéticos tienen lugar en etapas tempranas de la vida, incluido el periodo fetal, pueden afectar de manera permanente la función celular. Un ejemplo de ello es el tabaquismo materno durante el embarazo, ya que este se asocia a variaciones en la metilación del ADN en el feto. Estos cambios se relacionan con un incremento del riesgo de EPOC, así como de su gravedad [12].

1.4. Diagnóstico

La principal prueba diagnóstica contemplada en la guía GOLD (*Global Initiative for Chronic Obstructive Lung Disease*) para pacientes con sospecha de EPOC es la espirometría postbroncodilatadora [4], que proporciona el cociente entre el volumen espiratorio forzado en el primer segundo (FEV1) y la capacidad vital forzada (FVC). Esta prueba debe realizarse siempre en la fase estable de la enfermedad [16] y, si su valor es < 0.7, se demuestra la existencia de obstrucción bronquial.

No obstante, aunque la espirometría ofrece múltiples ventajas como alta reproducibilidad, bajo coste y facilidad de ejecución, tiende a infradiagnosticar en pacientes jóvenes (alcanzando cifras de hasta el 75%) [13] y a sobrediagnosticar en edades avanzadas. Esto se debe a que la especificidad del criterio FEV1/FVC < 0.7 disminuye con el envejecimiento [13]. Así, algunos pacientes fumadores sintomáticos pueden presentar una espirometría normal, mientras su parénquima pulmonar muestra alteraciones estructurales en una tomografía computarizada (TAC). Asimismo, los parámetros espirométricos pueden ser adecuados, pero la prueba de difusión de monóxido de carbono puede revelar daños funcionales [13].

Por consiguiente, el diagnóstico basado exclusivamente en la espirometría subestima el trastorno fisiológico del sujeto y pone de manifiesto la necesidad de criterios adicionales como la exposición a factores de riesgo (especialmente el humo de tabaco) y síntomas de carácter respiratorio como disnea de esfuerzo, tos con expectoración e infecciones respiratorias recurrentes en las vías aéreas bajas [16].

Por otra parte, existen pruebas complementarias al diagnóstico que permiten caracterizar la enfermedad, identificar comorbilidades y valorar el grado de afectación y pronóstico de la misma (Tabla 1) [13].

Tabla 1. Pruebas complementarias al diagnóstico.

Pruebas	Indicación
Analítica de sangre	Estimación del nivel de eosinófilos y α-1 antitripsina
Radiografía de tórax postero-anterior y lateral	Detección de atrapamiento aéreo e hiperinsuflación pulmonar
TAC	Valoración del enfisema y diagnóstico de bronquiectasias
Volúmenes pulmonares estáticos y prueba de difusión de monóxido de carbono	Detección de atrapamiento aéreo, hiperinsuflación e impacto funcional de enfisema
Test de la marcha de 6 minutos	Desaturación y capacidad de realizar ejercicio físico
Gasometría arterial	Sospecha de insuficiencia respiratoria y/o sospecha de hipercapnia
Cultivo de esputo	Descarte de infección bronquial en bronquitis crónica con exacerbaciones frecuentes
Ecocardiograma transtorácico	Sospecha de hipertensión arterial pulmonar y comorbilidad cardíaca
Estudio de sueño	Sospecha de apnea del sueño o hipertensión pulmonar desproporcionada al grado de obstrucción

1.5. Estratificación de pacientes

Tras la confirmación del diagnóstico de EPOC mediante espirometría y con el soporte de las pruebas complementarias pertinentes, es fundamental valorar la severidad de la enfermedad. Esta evaluación integral permite dictaminar el nivel de riesgo del paciente basándose en el grado de obstrucción bronquial, intensidad de los síntomas y número de exacerbaciones experimentadas en el último año [16]. Una correcta caracterización del riesgo no solo ayuda a

predecir las manifestaciones agudas de la enfermedad, sino que permite ofrecer opciones terapéuticas óptimas y personalizadas, adaptadas al perfil clínico del paciente en cuestión [16].

1.5.1. Gravedad de la obstrucción al flujo aéreo

Bajo la premisa de un cociente FEV1/FVC < 0.7, se establecen diferentes niveles de gravedad según el porcentaje del FEV1 con respecto al valor teórico obtenido en la espirometría postbroncodilatadora. De este modo, la guía GOLD clasifica la EPOC en cuatro grados (GOLD 1 a 4), donde una cifra mayor es sinónimo de un mayor nivel de severidad (Tabla 2). Cabe destacar que los puntos de corte espirométricos contemplados tienen como finalidad simplificar la clasificación clínica [4].

Tabla 2. Clasificación GOLD de la gravedad de la obstrucción al flujo aéreo.

Grado GOLD	Gravedad	FEV1 (% del esperado)
GOLD 1	Leve	≥ 80
GOLD 2	Moderada	50-79
GOLD 3	Grave	30-49
GOLD 4	Muy grave	<30

1.5.2. Clasificación según síntomas y exacerbaciones (GOLD A-B-E)

La guía GOLD propone una evaluación de los síntomas e historia de agudizaciones del año anterior con el objetivo de lograr una estratificación más precisa, yendo más allá de la caracterización de obstrucción al flujo aéreo.

Para el estudio de los síntomas clínicos, se emplean dos escalas validadas: la escala mMRC (*modified Medical Research Council*) y el test CAT (*COPD Assessment Test*).

- **Escala mMRC** Se trata del primer cuestionario desarrollado para cuantificar el grado de disnea, síntoma altamente prevalente en pacientes con EPOC [4]. Esta escala evalúa la dificultad respiratoria en relación con el nivel de esfuerzo físico aplicado. Así, la magnitud de la disnea en el paciente puede clasificarse atendiendo a su intensidad (Tabla 3), desde un grado 0 en el que la disnea está ausente excepto en esfuerzos intensos, hasta un grado 4 donde la gravedad es tan alta que incapacita al sujeto para cualquier actividad cotidiana básica [17].

Tabla 3. Escala de disnea mMRC.

ESCALA DE DISNEA MODIFICADA DEL BRITISH MEDICAL RESEARCH COUNCIL (mMRC)	
GRADO 0	Ausencia de disnea excepto al realizar ejercicio intenso.
GRADO 1	Disnea al andar deprisa en llano, o al andar subiendo una pendiente poco pronunciada.
GRADO 2	La disnea produce incapacidad para mantener el paso de otras personas de la misma edad caminando en llano o tener que parar a descansar al andar en llano al propio paso.
GRADO 3	La disnea produce la necesidad de tener que parar a descansar al andar unos 100 metros o después de pocos minutos de andar en llano.
GRADO 4	La disnea impide al paciente salir de casa o aparece con actividades como vestirse o desvestirse.

- **Test CAT.** Consiste en un cuestionario conformado por 8 ítems que revela el estado de salud en pacientes con EPOC. Cada una de las preguntas se califica de 0 (mínimo impacto en la calidad de vida) a 5 puntos (máximo impacto en el día a día del paciente). Por consiguiente, la puntuación total puede oscilar entre 0 y 40 puntos, y permite clasificar la repercusión de la enfermedad sobre el paciente [18]:
  - **Puntuación < 10:** bajo impacto en la calidad de vida del paciente.
  - **Puntuación 10-20:** impacto moderado.
  - **Puntuación > 20:** impacto alto.

  En la Figura 7 se presenta una plantilla de este test, mostrándose las distintas cuestiones por las que está conformado:

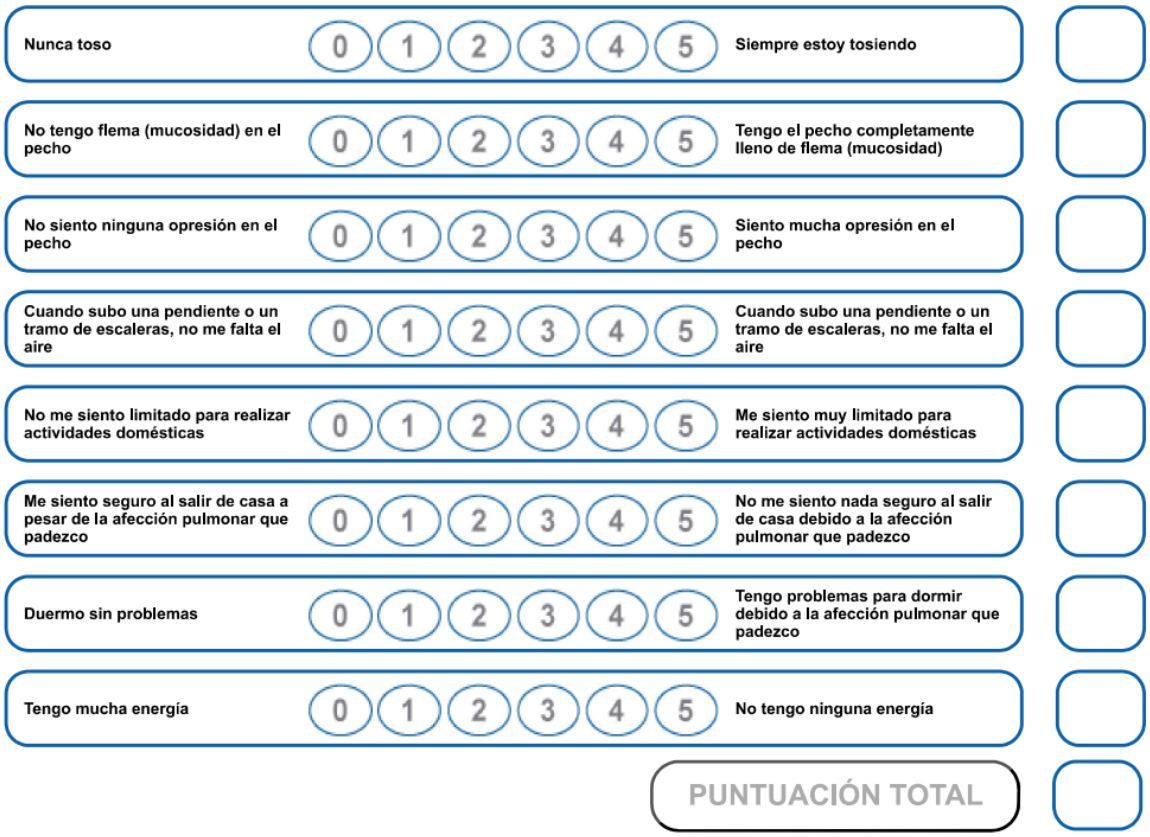

Figura 7. Plantilla del test CAT [19].

Además del marco sintomático, es importante considerar los antecedentes de las agudizaciones experimentadas. Para ello, debe cuantificarse el número de episodios en el último año, permitiendo estratificar al paciente en dos grupos [20]:

- ≥ 2 exacerbaciones moderadas o ≥1, requiriendo de ingreso hospitalario.
- 0 o 1 exacerbación sin necesidad de hospitalización.

Teniendo presente las consideraciones previas sobre sintomatología y número de exacerbaciones, la guía GOLD establece tres grupos: A, B y E (Tabla 4) [4]. El grupo E se caracteriza por una alta frecuencia de exacerbaciones, siendo la sintomatología variable a diferencia del resto de grupos [4].

Tabla 4. Caracterización GOLD 2025 de grupos A, B y E en EPOC.

	mMRC	CAT	Exacerbaciones
GRUPO A	0 - 1	< 10	0 -1 moderadas sin ingreso
GRUPO B	≥ 2	≥ 10	
GRUPO E	≥ 2 (típico) o 0-1	≥ 10 (típico) o < 10	≥ 2 moderadas o ≥ 1 con ingreso

Los sujetos del grupo A se caracterizan por presentar síntomas más leves y poco incapacitantes en la vida diaria, junto con pocas exacerbaciones que no motivan ingreso hospitalario. Por su parte, el grupo B presenta el mismo perfil de exacerbaciones, pero con mayor carga de síntomas e impacto en la calidad de vida. Finalmente, el grupo E se distingue por un elevado número de exacerbaciones o por la presencia de agudizaciones que precisan ingreso hospitalario, acompañado de sintomatología muy variable [4].

1.5.3. Clasificación del riesgo clínico según GesEPOC:

La Guía Española de la EPOC (GesEPOC) surge como una iniciativa de la Sociedad Española de Neumología y Cirugía Torácica (SEPAR), con la colaboración de sociedades científicas y del Foro Español de Pacientes. Su principal objetivo es mejorar la atención y calidad de vida de pacientes con EPOC, así como disminuir su prevalencia. Para ello, la guía integra los avances más recientes en diagnóstico, tratamiento y estratificación de la severidad, incluyendo además actualizaciones procedentes de la guía GOLD internacional [21].

En base a las categorizaciones abordadas previamente, la GesEPOC propone una segmentación del riesgo en bajo y alto [16]. Para determinar la inclusión en un grupo u otro, deben reunirse una serie de criterios (Figura 8).

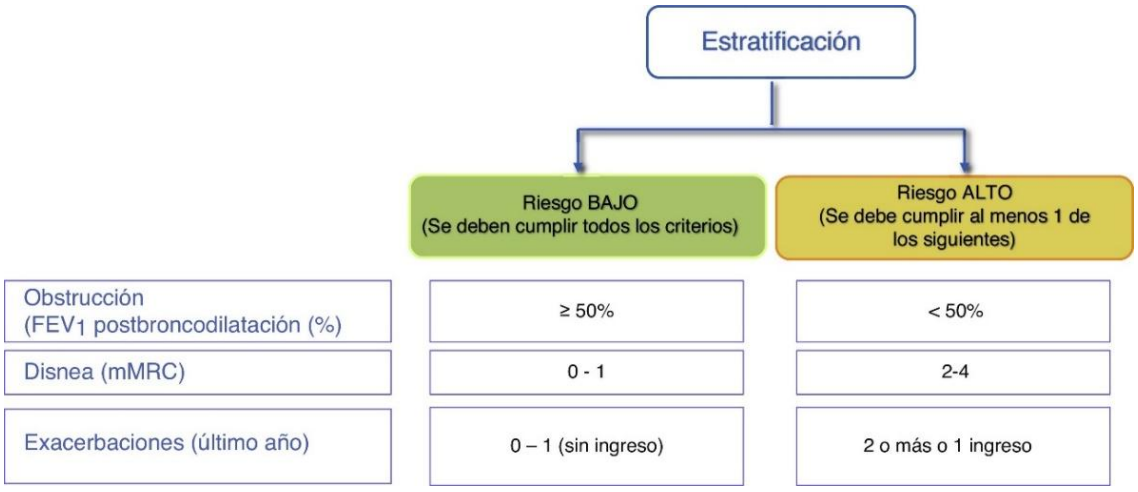

Figura 8. Estratificación del riesgo en EPOC [16].

Existen evidencias de que todos los componentes de esta estratificación permiten predecir la mortalidad. Además, investigaciones recientes han demostrado una fuerte asociación entre el nivel de riesgo, atención asistencial y selección del tratamiento óptimo [16]. Por tanto, esta categorización no solo orienta sobre el pronóstico, sino que sienta las bases para tomar las mejores decisiones terapéuticas atendiendo al perfil específico del paciente.

1.6. Tratamiento

Aunque la EPOC es una enfermedad crónica y progresiva y, por tanto, no curable, existen tratamientos farmacológicos y no farmacológicos cuyo objetivo es reducir los síntomas y minimizar los riesgos asociados [4]. De esta manera, se logra aliviar la sintomatología, mejorar la tolerancia al ejercicio físico y el estado de salud del paciente, así como prevenir la progresión de la enfermedad y el desarrollo de exacerbaciones, facilitando su manejo y contribuyendo a reducir la mortalidad [4].

1.6.1. Terapia propuesta en GOLD 2025

1.6.1.1. Opciones farmacológicas en EPOC estable

El tratamiento farmacológico se establece en función del grupo al que pertenece el paciente, determinado por el número de exacerbaciones, la carga sintomática y, en ciertos casos, el recuento de eosinófilos en sangre. Las indicaciones recogidas en la guía GOLD 2025 se resumen en la Tabla 5 [4]:

Tabla 5. Tratamiento farmacológico por grupos según GOLD 2025.

GRUPO A	GRUPO B	GRUPO E
Broncodilatadores	LABA + LAMA Monoterapia LABA/LAMA	LABA + LAMA LABA + LAMA +ICS

LABA: Long Acting Beta Agonists (agonistas β2 de acción prolongada).
LAMA: Long Acting Muscarinic Antragonists (antagonistas muscarínicos de acción prolongada).
ICS: corticoides inhalados.

- **Grupo A**: se recomienda el uso de broncodilatadores, que pueden ser de acción corta o prolongada según el grado de disnea del paciente. Ante una buena disponibilidad y reducido coste, es preferible el uso de broncodilatadores de acción duradera, salvo en casos de disnea ocasional [4].

- **Grupo B:** la terapia consiste en la combinación de LABA y LAMA, ambos broncodilatadores de acción prolongada. Su acción conjunta ha demostrado una eficacia superior al empleo exclusivo de LAMA. No obstante, si su aplicación no es viable, puede optarse por la monoterapia con LABA o LAMA. No existen evidencias que respalden la superioridad clínica de uno sobre otro, por lo que la selección debe basarse en la sensación de alivio del paciente en concreto [4].

  Cabe señalar la probabilidad de que un sujeto perteneciente a este grupo padezca comorbilidades que intensifiquen los síntomas y condicionen negativamente el pronóstico. Esto subraya la necesidad de identificarlas y tratarlas atendiendo a las guías nacionales e internacionales actuales.

- **Grupo E:** de forma análoga al grupo B, se opta por la aplicación conjunta de LABA y LAMA, dado que su actuación permite reducir las exacerbaciones en comparación a la monoterapia de broncodilatadores de acción prolongada. No obstante, si la hematología del paciente revela un nivel de eosinófilos en sangre ≥ 300 células µL, puede considerarse la triple terapia LABA+LAMA+ICS (corticoides inhalados). En caso de padecer asma además de EPOC, la aplicación de ICS es obligatoria [4].

Independientemente de la pertenencia a un grupo u otro, resulta fundamental revisar, evaluar y ajustar el tratamiento atendiendo al contexto específico del paciente [4]:

- Revisar: síntomas y posible riesgo de exacerbaciones.
- Evaluar: respuesta al tratamiento y técnica de inhalación.
- Ajustar: debe adaptarse el tratamiento atendiendo a la revisión clínica efectuada donde se estudian efectos adversos.

1.6.1.2. Opciones no farmacológicas:

Las intervenciones no farmacológicas complementan el tratamiento farmacológico. En todos los grupos, se considera prioritario el abandono del hábito tabáquico, recomendándose actividad física regular. Entre las terapias no farmacológicas pueden citarse: vacunas preventivas, rehabilitación pulmonar, oxigenoterapia domiciliaria y ventilación mecánica no invasiva (VNI) [4].

- **Vacunaciones preventivas:** Según las guías locales, se establece la administración de una serie de vacunas preventivas para disminuir el riesgo de padecer infecciones respiratorias que puedan agravar la EPOC. Estas vacunaciones incluyen: influenza, COVID-19, neumococo, tos ferina, herpes zóster y virus respiratorio sincitial (RSV) [4].
- **Rehabilitación pulmonar:** Resulta clave para los grupos B y E. Esta terapia se adapta al perfil concreto del paciente y a sus comorbilidades. Se recomienda realizar ejercicio supervisado al menos dos veces por semana, y sus beneficios son observables tras programas de entre 6 y 8 semanas de duración [4].

- **Oxigenoterapia domiciliaria:** Permite mantener una saturación de oxígeno en sangre arterial (SaO_2) ≥ 90% y está indicada en dos casos [4]:

    - Pacientes con una presión de oxígeno arterial (PaO_2) ≤ 55 mmHg o una saturación de oxígeno ≤ 88%, en presencia o no de hipercapnia.
    - Pacientes con PaO2 entre 55-60 mmHg o una saturación del 88% en caso de padecer también hipertensión pulmonar, insuficiencia cardíaca congestiva o policitemia.

  Tras un periodo de 60-90 días, es necesario reevaluar si la oxigenoterapia sigue siendo necesaria y si su aplicación ha resultado efectiva [4].

- **Ventilación mecánica no invasiva (VNI):** Se aplica en sujetos con EPOC estable muy severa, especialmente en aquellos que padezcan hipercapnia diurna significativa e ingreso hospitalario reciente. En aquellos casos en los que se tenga EPOC y apnea obstructiva del sueño (AOS), se han observado mejoras gracias a la aplicación de CPAP (presión positiva continua en la vía aérea). Además, la VNI permite reducir la mortalidad, la morbilidad en exacerbaciones agudas con presencia de insuficiencia respiratoria y el riesgo de hospitalización [4].

En algunos estudios, se ha demostrado que el uso conjunto de oxigenoterapia y VNI en el domicilio retrasa el tiempo hasta una nueva hospitalización o incluso el fallecimiento en los 12 meses siguientes [4].

Si la respuesta a estos tratamientos es adecuada, deben ofrecerse vacunas anuales, educación para que los pacientes sean capaces de manejar la enfermedad en la medida de lo posible y valoración de los factores conductuales (exposición a factores de riesgo). En caso de que la respuesta no resulte óptima, deben abordarse los rasgos tratables como disnea o exacerbaciones [4].

1.6.2. Terapia propuesta por la GesEPOC

La GesEPOC establece las pautas terapéuticas en función del fenotipo que presenta el paciente. Así, pueden citarse tres fenotipos posibles: no agudizador, agudizador eosinofílico y agudizador no eosinofílico [16].

- **No agudizador**: pacientes que han experimentado, como máximo, una agudización leve en el último año sin precisar ingreso hospitalario. Para este perfil clínico, se aplica doble broncodilatación (LABA + LAMA), ya que se ha evidenciado una mayor eficacia frente a la monoterapia. Las mejorías se reflejan en la disminución de disnea, una mejor calidad de vida y la reducción de la necesidad de administrar medicación de rescate [16].

- **Agudizador eosinofílico**: un paciente agudizador es aquel que ha sufrido dos o más agudizaciones ambulatorias, o al menos una que haya requerido hospitalización el año previo. Para distinguir si estos eventos se deben a una recaída o al fracaso del tratamiento, debe transcurrir un intervalo temporal mínimo de cuatro semanas desde la resolución de la agudización anterior, o seis semanas desde que comenzó la sintomatología. Por otra parte, el fenotipo eosinofílico se define por la presencia de > 300 eosinófilos/mm^3 en sangre periférica en fase estable. En este perfil de pacientes, la mayor respuesta al tratamiento se obtiene con la combinación de ICS y LABA, dado que contribuye a un menor riesgo de exacerbaciones. No obstante, investigaciones recientes han concluido que la triple terapia (LABA + LAMA + ICS) brinda mejorías más significativas en la función pulmonar, en los

síntomas respiratorios y en la disminución de las exacerbaciones en comparación con la aplicación de ICS y LABA [16].

- **Agudizador no eosinofílico:** pacientes cuyo perfil encaja con las características del fenotipo agudizador, pero que poseen un número de eosinófilos < 300 células/mm^3 en sangre periférica. Los ICS presentan una eficacia limitada en estos casos, aunque pueden administrarse especialmente si los eosinófilos superan las 100 células/mm^3. Si el recuento es < 100 células/mm^3, se desaconseja su empleo debido al riesgo de efectos secundarios. En pacientes con cifras de 100-300 células/mm^3, se valora la triple terapia (LABA + LAMA + ICS). Sin embargo, en la mayoría de los casos se aconseja la aplicación de LABA y LAMA. Esta combinación resulta más eficaz que LABA y ICS cuando las exacerbaciones son menos frecuentes o se requieren antibióticos [16].

1.7. Exacerbaciones de EPOC

1.7.1. Concepto e implicaciones de las exacerbaciones

La Guía GOLD 2025 define la exacerbación de la EPOC como un evento caracterizado por el incremento de alguno de los siguientes síntomas: disnea, tos o esputo, con un empeoramiento que ocurre en menos de 14 días [4]. Es frecuente que concurran otros signos clínicos, como taquipnea y/o taquicardia, así como inflamación local y sistémica desencadenada por infecciones, exposición a contaminantes ambientales o irritantes [4].

No obstante, no existe una definición universal sobre “exacerbación de la EPOC”, lo que provoca que a menudo se infravaloren [22]. Esto constituye un obstáculo relevante, ya que, al ser una enfermedad crónica y progresiva, muchos sujetos no solicitan de forma oportuna atención médica por atribución de los síntomas a causas benignas como el envejecimiento [22]. Esto implica una detección en fases más tardías, lo que agrava el impacto de la EPOC sobre el paciente.

A pesar de esta falta de consenso, las exacerbaciones son ampliamente reconocidas como elementos clave con repercusiones significativamente negativas en la historia natural de la EPOC. Sus consecuencias incluyen un empeoramiento de la función pulmonar, deterioro en la calidad de vida y aumento de la morbilidad [20]. En la Figura 9 se recogen otros efectos importantes.

Figura 9. Efectos de las exacerbaciones [20].

TVP: trombosis venosa profunda; TEP: tromboembolismo pulmonar; IAM: infarto agudo de miocardio; ACV: accidente cerebrovascular.

1.7.2. Diagnóstico y clasificación de la severidad de las exacerbaciones de EPOC

El diagnóstico diferencial en las agudizaciones es esencial para poder distinguir entre patologías. Las pruebas efectuadas con este fin más comunes son [22]:

- **Radiografía de tórax:** detecta neumonía, neumotórax, derrame pleural o insuficiencia cardíaca congestiva.
- **Determinación de dímero D y angio-TC torácico**, por sospecha de tromboembolia pulmonar.
- **Electrocardiograma y determinación de troponinas:** orientados a identificar arritmias o síndrome coronario agudo.

Actualmente no se dispone de un biomarcador diagnóstico único que presente una elevada especificidad y sensibilidad. Por ello, el diagnóstico se efectúa basándose en tres criterios objetivables: agravamiento de la disnea, desaturación de oxígeno y biomarcadores alterados [22].

Tras establecer el diagnóstico, es importante categorizar la gravedad de la exacerbación para adoptar las intervenciones terapéuticas más adecuadas. En este sentido, la Propuesta Roma

presenta una clasificación de la gravedad de las agudizaciones en tres grados: leve, moderada y grave [22]. La inclusión en uno u otro se basa en seis parámetros: disnea, frecuencia respiratoria, frecuencia cardíaca, saturación de oxígeno, proteína C reactiva (PCR) y valores gasométricos. En la Tabla 6 se muestra esta clasificación [22]:

Tabla 6. Grados de severidad de las exacerbaciones de EPOC (Impacto de las exacerbaciones en la enfermedad pulmonar obstructiva crónica) [22].

SEVERIDAD	CRITERIOS
LEVE	➢ Disnea < 5 evaluada por la escala análoga visual (VAS). ➢ Frecuencia respiratoria < 24 respiraciones/min. ➢ Frecuencia cardíaca < 95 latidos/min. ➢ SaO_2 en reposo > 92%, ya sea respirando aire ambiente o con oxigenoterapia, y/o variación ≤ 3% respecto a su valor basal de SaO_2. ➢ PCR < 10 mg/L.
MODERADA (Cumplir al menos 3 de los criterios expuestos)	➢ Disnea ≥ 5 evaluada por VAS. ➢ Frecuencia respiratoria ≥ 24 respiraciones/min. ➢ SaO_2 basal < 92% respirando aire ambiente (o con oxigenoterapia habitual) y/o variación de la SaO_2 ≤ 3% respecto a su valor en reposo. ➢ PCR ≥ 10 mg/L. ➢ Hipoxemia (PaO_2 ≤ 60 mmHg) y/o hipercapnia ($PaCO_2$ > 45 mmHg), pero sin acidosis (pH > 7.35).
GRAVE	➢ Hipercapnia con acidosis (pH <7.35).

VAS: escala análoga visual; SaO_2: saturación de oxígeno en sangre arterial; PCR: proteína C reactiva: PaO_2: presión de oxígeno en sangre arterial; $PaCO_2$: presión de dióxido de carbono en sangre arterial.

1.7.3. Reingresos y mortalidad por exacerbaciones

Tras una exacerbación moderada o grave que requiere hospitalización, el paciente presenta una mayor probabilidad de reingreso a los 30 y 90 días del alta. Este riesgo aumenta si el sujeto posee antecedentes de agudizaciones previas que hayan precisado hospitalizaciones prolongadas. Los reingresos frecuentes por exacerbaciones de EPOC se asocian con una mayor mortalidad [22].

Aunque la cantidad de reingresos depende del contexto demográfico, en el meta-análisis de Ruan et al. (2023) [23], en el que se analizan un total de 46 estudios de distintos territorios, se recogieron las tasas de reingreso por exacerbación 30, 60, 90, 180 y 365 días después del alta. De este modo, se ofrecen tasas de reingreso globales más precisas y generalizables que las aportadas por estudios individuales cuyos porcentajes pueden depender del contexto nacional. A los 30 días, la tasa promedio combinada anual fue del 11%, a los 60 días del 17%, a los 90 días del 30% y a los 180 días del 37% [23].

En cuanto a la mortalidad intrahospitalaria, se han identificado diversos factores que contribuyen a su aumento, como disfunciones cardíacas, estancias prolongadas, envejecimiento, comorbilidades, desnutrición y valores alterados de la gasometría arterial al ingreso [22].

Por otra parte, la mortalidad posterior a la hospitalización es especialmente alta durante la primera semana tras el alta y se mantiene elevada tres meses después. Además, es importante señalar que la tasa de supervivencia a cinco años después de un ingreso por exacerbación de EPOC es inferior al 50% [22].

1.7.4. Factores de riesgo de reingreso por exacerbación

Las exacerbaciones de la EPOC constituyen la principal carga sanitaria de la enfermedad [4]. Por ello, la identificación de los factores de riesgo de reingreso por exacerbaciones de la EPOC resulta fundamental para poder implementar estrategias preventivas, mejorar la comprensión de la enfermedad y reducir su impacto. En este sentido, a continuación, se especifican algunos de los factores predisponentes [20], [22], [24]:

- **Infecciones**. Las exacerbaciones se deben en su mayoría a causas infecciosas. El 29.7% de las mismas son de causa bacteriana, el 23.3% se deben a virus y el 25% se corresponden a coinfecciones virales/bacterianas.
- **Antecedentes de exacerbaciones**, sobre todo en el último año.
- **Tabaquismo.**
- **Presencia de comorbilidades.**
- **Uso de oxigenoterapia.**
- **Grado de severidad**. Existe una clara relación entre el grupo GOLD y la frecuencia de las exacerbaciones. A mayor grado GOLD, mayor predisposición.

Conocer estos factores no solo permite optimizar el manejo clínico y disminuir la carga socioeconómica y sanitaria, sino que resulta muy relevante para seleccionar las variables con las que elaborar los modelos predictivos como el del presente trabajo.

1.7.5. Medidas preventivas y manejo de las exacerbaciones

La frecuencia de las exacerbaciones puede disminuirse mediante diversas intervenciones según la Guía GOLD 2025. Entre ellas, se incluyen [4]:

- **Broncodilatadores**: LABA, LAMA o combinación de ambas.
- **Corticosteroides**: LABA + ICS o LABA + LAMA + ICS.
- **Antiinflamatorios no esteroideos**.
- **Antiinfecciosos** como vacunas.
- **Agentes mucolíticos**.

- **Otros**: abandono del hábito tabáquico, rehabilitación, vitamina D y medidas de protección como uso de mascarilla o higiene de manos.

Por otra parte, el manejo de las agudizaciones se fundamenta en cinco pilares [20]:

- **Broncodilatadores:** empleo de SABA (agonistas β2 de corta acción inhalados), con posibilidad de combinación con SAMA (anticolinérgico de corta acción). Estos son administrados mediante un inhalador o nebulizador.

- **Glucocorticoides:** optimizan la función pulmonar, mejoran la oxigenación y evitan recaídas tempranas. Esto contribuye a prevenir el fracaso de la terapia y reducir la estancia hospitalaria. Sin embargo, el uso prolongado de estos fármacos puede favorecer el desarrollo de neumonía e incrementar la mortalidad.

- **Antibióticos**: indicados en pacientes con aumento del volumen de esputo y aspecto purulento del mismo. La elección del antibiótico depende de los patrones de resistencia bacteriana del sujeto y de las guías locales. El periodo de administración es de aproximadamente 5-7 días.

- **Oxígeno:** aplicado en situaciones de hipoxemia, con una saturación objetivo del 88-92% y con el fin de prevenir el desarrollo de hipercapnia.

- **Ventilación no invasiva:** recomendada en casos de acidosis respiratoria (pH 7.2 -7.35). Permite mejorar el intercambio gaseoso y disminuir la frecuencia respiratoria, el trabajo respiratorio, la severidad de la disnea, la tasa de intubación, la estancia hospitalaria y la mortalidad. En este contexto, la cánula nasal de alto flujo (CNAF) se ha ido consolidando recientemente como una herramienta eficaz para tratar la exacerbación.

1.7.6. Volumen global de las exacerbaciones

El análisis en este apartado se basa en las proyecciones recogidas en un estudio reciente de Boers et al. 2025 [25]. En este artículo se emplea un modelo de simulación desarrollado a partir de datos objetivos como prevalencia, incidencia y factores de riesgo extraídos de fuentes como *la Global Burden of Disease (GBD)* y el *World Bank*. Con ello, se efectúan estimaciones del número de exacerbaciones desde el presente (2025) al año 2050. Estos datos se recogen en la Tabla 7 [25], donde se muestra un desglose de la cantidad estimada de exacerbaciones de EPOC por región, así como la variación porcentual de agudizaciones para los años considerados.

Tabla 7. Número de exacerbaciones de EPOC anuales estimadas [25].

NÚMERO DE EXACERBACIONES ANUALES (en millones)			
Región	**2025**	**2050**	**Variación porcentual (%)**
Asia Oriental y el Pacífico	171.30	236.90	**+38.30%**
Asia del Sur	145.40	193.60	**+33.10%**
Oriente Medio	15.10	19.10	**+26.50%**
Latinoamérica	6.60	7.40	**+12.10%**
Europa	49.10	48.10	**-2.00%**
Recuento global	**456.40**	**567.60**	**+24.30%**

La fuente de la que se obtiene la tabla incluye también datos de Norteamérica. Sin embargo, se han observado una inconsistencia en los mismos y se ha optado por no incluirlos para no inducir errores.

Puede apreciarse un aumento generalizado del volumen de exacerbaciones a nivel mundial, con un incremento estimado del 24.30% entre 2025 y 2050. Todas las regiones experimentan un crecimiento, a excepción de Europa, donde se proyecta una caída leve (-2.0%). El mayor aumento se corresponde con Asia Oriental y el Pacífico (+38.3%), seguido de cerca por Asia del Sur (+33.1%), reflejando ambas cargas relevantes. Otros territorios como Oriente Medio y Latinoamérica, a pesar de registrar cifras absolutas inferiores al resto, también presentan crecimientos importantes.

Ante esta tendencia global de las exacerbaciones, se pone de manifiesto la necesidad de confeccionar herramientas capaces de anticipar eventos que puedan suponer una complicación tanto para el paciente como para el sistema sanitario en materia de recursos. En este sentido, los modelos predictivos resultan clave para ofrecer cuidados personalizados y diseñar estrategias eficientes que contribuyan a disminuir el impacto socioeconómico de la EPOC.

1.8. Impacto socioeconómico de la EPOC y sus exacerbaciones

A pesar de que la EPOC representa una crisis significativa de salud pública, continúa siendo infradiagnosticada y no recibe la atención necesaria. Se estima que cada hora fallecen 425 personas en todo el mundo por esta patología [26], cobrándose la vida de más personas que el cáncer de pulmón y de mama combinados. En el año 2024, afectaba a 391 millones de personas a nivel global [26].

Su devastadora carga subraya la necesidad de otorgar a la EPOC la prioridad adecuada, así como una financiación suficiente y opciones terapéuticas óptimas. Las consecuencias de la enfermedad abarcan tanto el ámbito social como el económico. La EPOC reduce notablemente la calidad de vida de los pacientes y de su entorno familiar, al tiempo que representa una elevada carga para el sistema sanitario, que requiere inversiones considerables de recursos para hacerle frente.

1.8.1. Carga económica

Costes totales asociados a la EPOC

Las muertes asociadas a esta enfermedad disminuyen la población en edad laboral y repercuten negativamente en la productividad. Asimismo, se observa un incremento del absentismo laboral [27], ya que el 40% de las personas con EPOC tienen que reducir su jornada o incluso abandonar su empleo [26]. En países carentes de un sistema de sanidad pública, la economía del núcleo familiar se ve condicionada, puesto que deben asumir los gastos derivados de las terapias. Puede observarse el mismo efecto en las aseguradoras, ya que, para cubrir los costes y garantizar una atención plena, deben incrementar necesariamente las primas. Ambos factores contribuyen a impedir la acumulación de capital físico en la economía [27].

El efecto que produce la EPOC depende del contexto económico de cada nación. Los países con ingresos bajos y medianos (PIBM) son los más afectados, puesto que sus habitantes enfrentan múltiples dificultades socioeconómicas y la atención médica es, en muchos casos, limitada. De hecho, la OMS afirma que alrededor del 90% de decesos por EPOC en personas menores de 70 años se producen en este tipo de países. Además, aproximadamente la mitad de los pacientes con esta patología en estas regiones experimenta exacerbaciones que requieren la aplicación de terapias o incluso motivan la hospitalización [26].

En los 365 días tras estas agudizaciones, se estima que el coste medio por paciente en urgencias u hospitalizado supera los 6000 euros. Los mayores gastos se dan en la hospitalización, pero también en el reingreso, puesto que normalmente un tercio de los pacientes que han sido dados de alta requieren ingresar de nuevo meses después [28].
Estudios efectuados por la OMS junto con otras entidades, evidencian que el acceso a la terapia inhalada en los PIBM es muy limitado y que el coste de los medicamentos resulta prácticamente

inasumible. El motivo de esto es que, gran parte de los fármacos disponibles proceden de marcas prestigiosas y existen pocas alternativas más asequibles [4]. A esto hay que añadir que los ministerios de salud no han confeccionado políticas sanitarias suficientes encaminadas a la prevención de la EPOC [27].

En la Unión Europea (UE) los costes totales destinados a las enfermedades respiratorias representan el 6% del presupuesto anual de salud. De esta cuantía, la EPOC abarca el 56%, lo que se traduce en un total de 38 600 millones de euros [4].

La confección de proyecciones de la carga económica de la EPOC y cómo se distribuye la misma entre países es imprescindible para tomar las medidas necesarias que permitan disminuir la morbilidad y mortalidad de la EPOC. En este sentido, se ha observado un impacto económico mayor en los próximos años. En Estados Unidos, a fecha de 2038, se espera que los costes asociados a EPOC supongan 800 900 millones de dólares (aproximadamente 40 000 millones de dólares anuales) [25]. En otras regiones como Inglaterra y Escocia, se espera alcanzar un coste directo total de 2900 millones de dólares y 264 millones de dólares, respectivamente [25].

En la Figura 10 [25], se muestra un mapa mundial procedente del estudio de Boers et al. (2025) [25], donde se representa la variación porcentual de los costes directos anuales por EPOC entre 2025 y 2050.

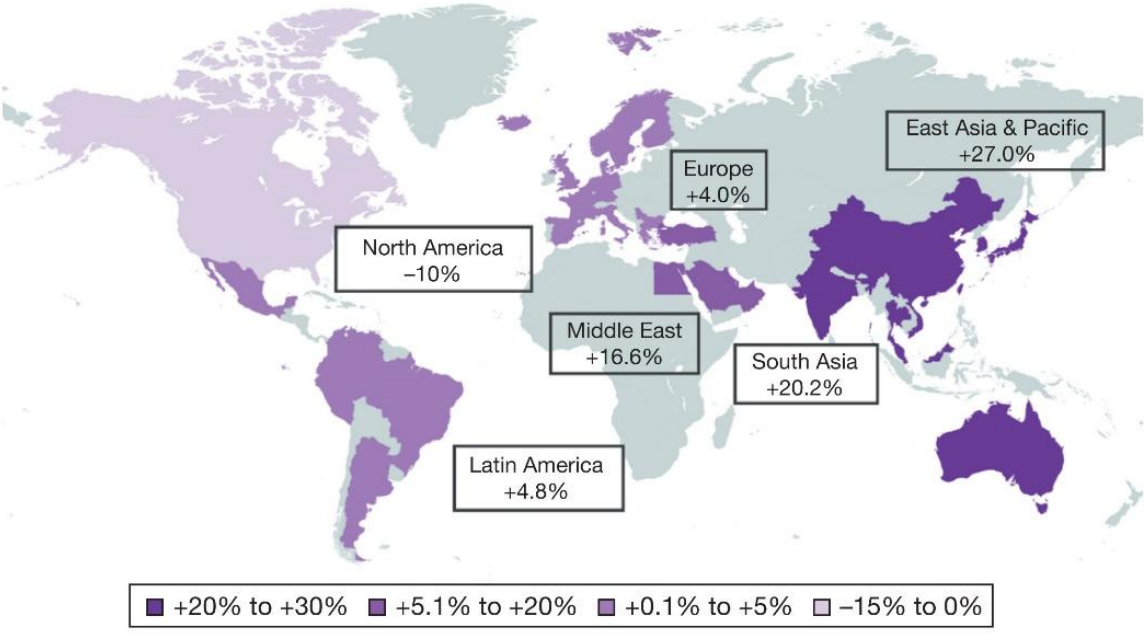

Figura 10. Mapa mundial de la variación porcentual estimada de los costes directos anuales asociados a EPOC entre 2025 y 2050 [25].

Lo primero que puede observarse es que la evolución de estos costes es totalmente desigual entre regiones. El incremento más destacado se registra en Asia Oriental y el Pacífico (+27%), que incluye naciones como Japón, Corea del Sur y Filipinas, así como territorios del Pacífico como Australia. Le sigue Asia del Sur, con países como Pakistán, India y Bangladés. Europa es la zona con menor incremento estimado (+4%) y Norteamérica es la única cuya variación porcentual es negativa, aunque no en gran medida (-10%).

Para complementar la información expuesta, se presenta la Tabla 8, procedente del mismo estudio [25], donde se detallan los costes directos anuales asociados a la EPOC para cada región. La tabla comprende cuatro columnas diferentes: región, costes estimados en 2025, costes predichos para 2050 y variación porcentual entre ambas fechas. Las regiones han sido ordenadas de mayor a menor variación porcentual para una mejor compresión.

Tabla 8. Costes directos anuales por EPOC desde 2025 hasta 2050 por cada región [25].

COSTES DIRECTOS ANUALES			
Región	**2025**	**2050**	**Variación porcentual (%)**
Asia Oriental y el Pacífico	305.49 mil millones USD	388.49 mil millones USD	**+27.0%**
Asia del Sur	52.70 mil millones USD	63.34 mil millones USD	**+20.2%**
Oriente Medio	22.80 mil millones USD	26.58 mil millones USD	**+16.6%**
Latinoamérica	15.66 mil millones USD	16.44 mil millones USD	**+4.8%**
Europa	170.01 mil millones USD	176.87 mil millones USD	**+4.0%**
Norteamérica	211.99 mil millones USD	190.69 mil millones USD	**-10.0%**
Coste global	**778.66 mil millones USD**	**862.42 mil millones USD**	**+10.8%**

Como puede apreciarse, se ha incluido una fila de coste global que resulta del sumatorio total de los datos de las columnas. Se estima un crecimiento mundial de costes directos por EPOC entre 2025 y 2050 de +10.8%. No obstante, pueden observarse detalles que en la Figura 10 eran imperceptibles.

Algunas regiones presentan costes elevados con variaciones en el periodo temporal considerado leves o incluso reducciones. Esto ocurre en el caso de Norteamérica, que ocupa la segunda posición en cuanto a costes más altos, pero presenta un decrecimiento porcentual. Sin embargo, otros territorios con una menor carga económica absoluta sufren aumentos significativos, como

es el caso de Asia del Sur y Oriente Medio. Por tanto, no existe una relación entre la cuantía del gasto y el porcentaje de variación estimado. Esto resalta la importancia de tener en cuenta los valores relativos y absolutos para así tener una comprensión correcta del impacto económico real de la EPOC.

En el contexto europeo, España también enfrenta una carga económica importante. En 2020, la Estrategia en EPOC del Sistema Nacional de Salud, elaborada por el Ministerio de Sanidad, estimó un coste anual de entre 750 y 1000 millones de euros [10]. Esta cuantía incluye costes directos, indirectos e intangibles. Por cada paciente, el coste medio directo oscila entre 1712 y 3238 euros anuales. Dentro de estos costes, se distinguen tres aspectos principales: hospitalización (40-45%), tratamientos (35-40%) y consultas y pruebas diagnósticas (15-25%) [10].

Según el Informe Anual del Sistema Nacional de Salud de 2023, los pacientes con EPOC requieren 2.5 veces más atención primaria que el resto de la población [29]. Cada año se registran 1.5 hospitalizaciones por cada 1000 habitantes, con una estancia media de 8 días. De estos ingresos, tres de cada cuatro corresponden a varones. Además, el número de urgencias atendidas por esta patología es aproximadamente de 90 000 [29].

Costes asociados a exacerbaciones

En la Figura 11 se muestra una estimación de los costes derivados de las exacerbaciones producidas anualmente en un intervalo comprendido entre 2025 y 2050 a nivel global [25]. En el año 2025, el gasto estimado es de 456 millones de dólares, mientras que en 2050 la cifra asciende hasta alcanzar los 568 millones de dólares. Esto se traduce en un incremento de costes del 24.6% en los 25 años considerados.

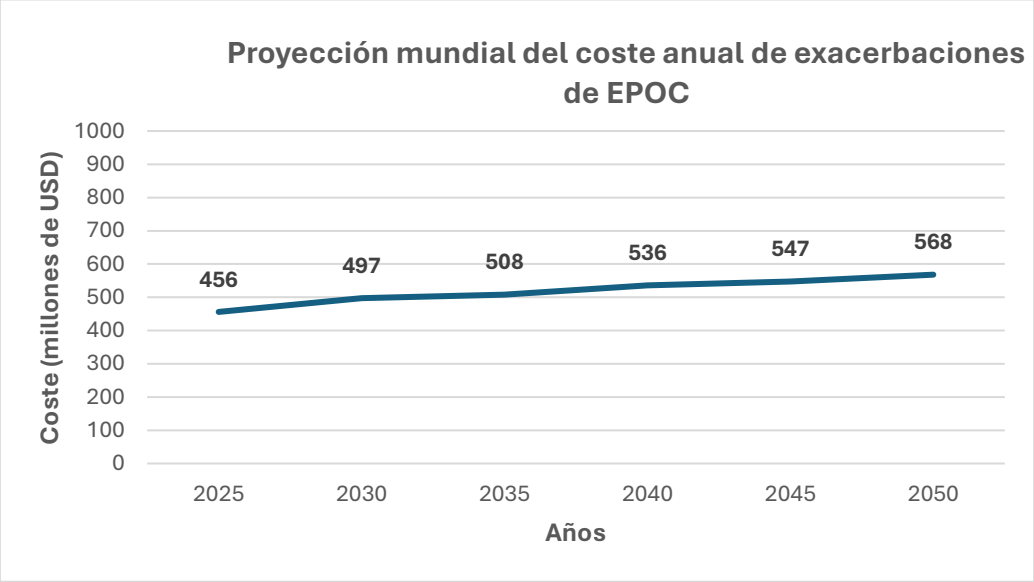

Figura 11. Proyección del coste anual de exacerbaciones de EPOC entre 2025 y 2050 a nivel mundial. Elaboración propia. Datos extraídos de [25].

Aunque la tendencia de crecimiento observada no es abrupta, refleja una evolución progresiva hacia cifras cada vez más altas. En consecuencia, la carga económica provocada por las exacerbaciones podría intensificarse con el tiempo. Un manejo correcto de los reingresos por exacerbaciones mediante modelos predictivos puede contribuir a mitigar este impacto.

1.8.2. Carga social

La carga social de la EPOC se manifiesta principalmente en la mortalidad y la discapacidad asociadas a esta afección. En este contexto, el estudio de la Carga Global de Enfermedad (*Global Burden of Disease,* GBD) emplea para su estimación los Años de Vida Ajustados por Discapacidad (AVAD, o DALY en inglés) [4]. El cálculo de los AVAD para una patología concreta se basa en la suma de los años de vida perdidos (AVP) y los años vividos con discapacidad (AVD), ponderados según la gravedad de la condición clínica [30].

Entre 1990 y 2019, la EPOC fue la responsable predominante de un incremento global de los DALY, con un impacto más acentuado en los PIBM. Durante ese mismo periodo, la carga mundial asociada a esta enfermedad experimentó un crecimiento desde los 59.2 millones de DALY en 1990 a 74.4 millones en 2019. Esto supone un crecimiento del 25.7%, reflejado especialmente en el sudeste asiático, India, África subsahariana y Sudamérica [4].

En países de ingresos elevados como Estados Unidos, el impacto es aún mayor, ya que en la actualidad la EPOC constituye la segunda causa de pérdida de DALY, siendo la primera la enfermedad cardíaca isquémica [4].

1.9. Inteligencia Artificial y EPOC

1.9.1. Introducción a la Inteligencia Artificial (IA)

El origen teórico de la informática se remonta a 1936, cuando el matemático británico Alan Turing presentó la noción de la Máquina de Turing (MT). Este modelo teórico permitía la manipulación de símbolos en una cinta dividida en celdas siguiendo un conjunto de reglas [31], lo que sentó las bases fundamentales de la computación [32].

En 1941, el ingeniero alemán Konrad Zuse desarrolló la Z3 [33], considerada la primera computadora programable. Esta máquina operaba con un sistema binario y podía almacenar un máximo de 64 palabras de 22 bits [32], [33]. Este avance, junto con la Máquina de Turing, constituyeron dos hitos que potenciaron el desarrollo de sistemas más complejos.

En 1956 se celebró la Conferencia de Dartmouth, un encuentro interdisciplinar que reunió a especialistas en matemáticas, neurología, psicología e ingeniería eléctrica. Fruto de esta reunión

surgió oficialmente el término de "Inteligencia Artificial", acuñado por el matemático John McCarthy. Por ello, este evento marcó el nacimiento formal de la IA [34], considerando esta disciplina como un nuevo campo de investigación.

Desde su reconocimiento oficial, la Inteligencia Artificial ha experimentado una evolución importante a lo largo de las últimas décadas, como se ilustra en la Figura 12 [35]:

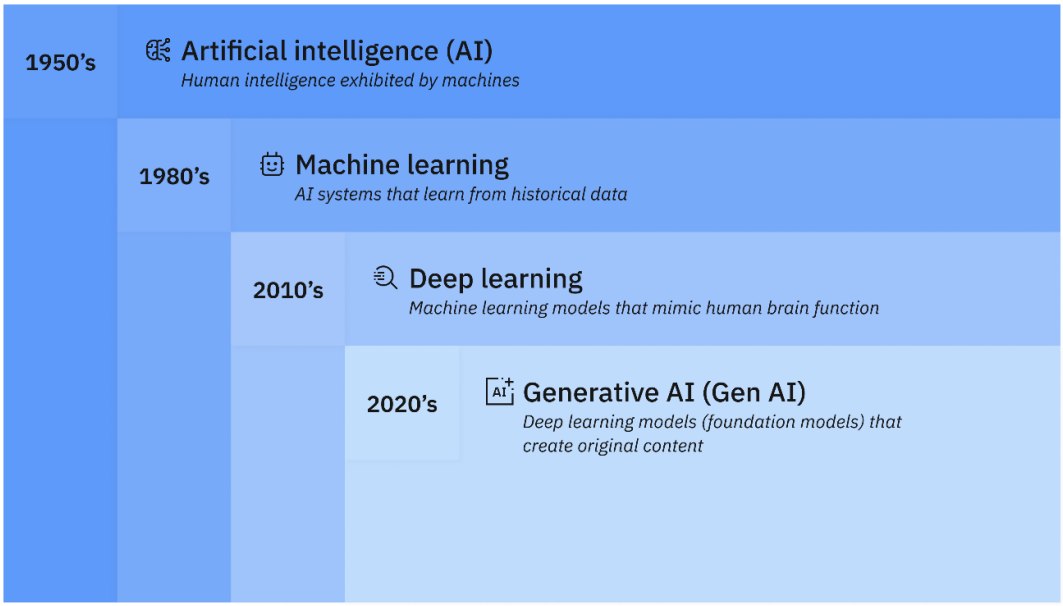

Figura 12. Evolución histórica de la Inteligencia Artificial (IA) en las últimas décadas [35].

- Década de 1950: Inteligencia Artificial. Etapa inicial en la que surge el concepto general de IA, concebida como la capacidad de las máquinas para reproducir la lógica humana.
- Década de 1980: *Machine Learning (ML).* Desarrollo de sistemas de IA capaces de aprender a partir de datos históricos. Aplicando algoritmos complejos es posible analizar grandes bases de datos, identificar patrones y generar predicciones. Su rendimiento mejora notoriamente a medida que aumenta el volumen de datos disponible en el entrenamiento [36].
- Década de 2010: *Deep Learning (DL).* Aparición de modelos de ML basados en redes neuronales profundas (multicapa), construidas para imitar de manera simplificada el funcionamiento del cerebro humano [37].
- Década de 2020: *Generative AI (Gen AI).* Transición de los modelos de DL capaces de realizar análisis de datos hacia sistemas con la habilidad de crear contenido original en respuesta a instrucciones del usuario [38].

En la actualidad, la Inteligencia Artificial se concibe como un conjunto de algoritmos que permite a los sistemas informáticos analizar datos y tomar decisiones en función de su conocimiento

adquirido. Las tareas que puede ejecutar son propias de la inteligencia humana, como identificar patrones, comprender lenguaje natural y resolver problemas [39].

1.9.2. Presencia de la Inteligencia Artificial en la EPOC

El impacto social de la Inteligencia Artificial es innegable y su presencia ha transformado prácticamente todos los ámbitos de la vida diaria [40]. El sector sanitario no es una excepción, ya que la IA se ha integrado progresivamente en el entorno de los sistemas de salud. Su aplicación resulta cada vez más relevante, puesto que impulsa mejoras en la productividad, la calidad asistencial, el acceso a los servicios y la participación activa del paciente en el manejo y seguimiento de su patología. Todo ello se traduce en una mayor humanización de los procesos, al favorecer una mayor dedicación de tiempo de calidad a la relación médico-paciente [41].

En la EPOC en concreto, la aplicación de la IA ha experimentado un crecimiento sostenido en los últimos años. Esto se evidencia en el incremento del número de publicaciones en buscadores científicos de gran reconocimiento como PubMed y ScienceDirect sobre esta temática. Los resultados anuales proporcionados por ambas plataformas bajo los términos de búsqueda "*artificial intelligence COPD*" reflejan un aumento importante a lo largo del tiempo, sobre todo a partir de 2018 como puede verse en las Figuras 13 y 14.

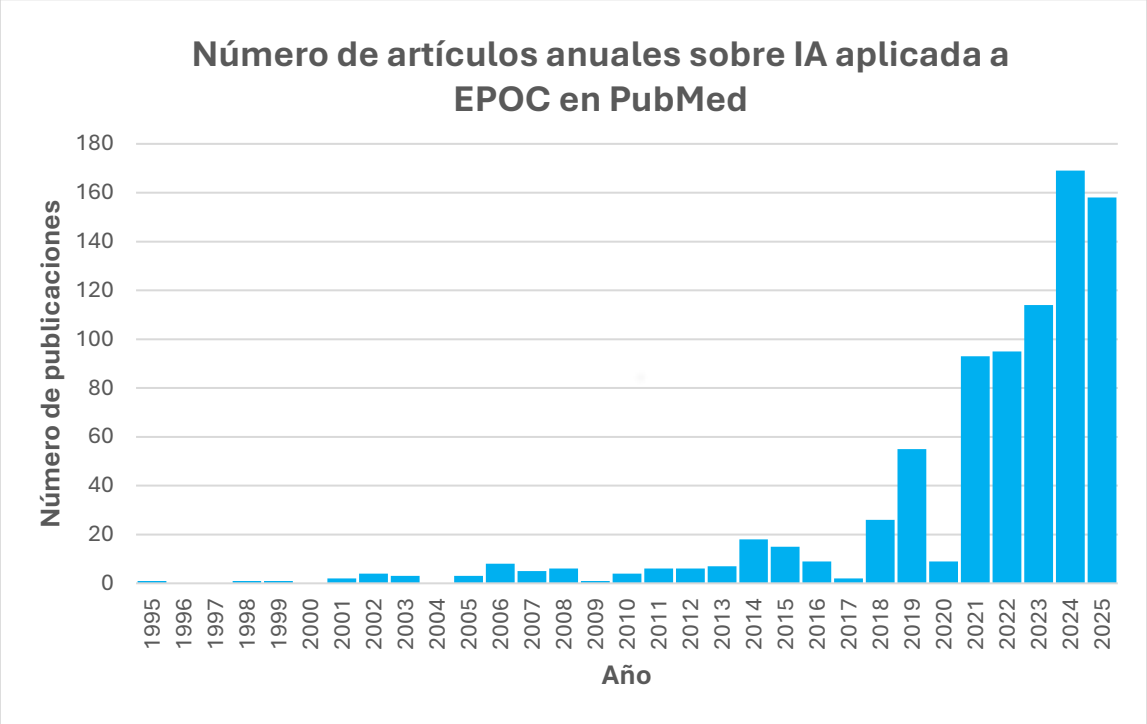

Figura 13. Número de publicaciones en PubMed sobre Inteligencia Artificial en EPOC entre 1995 y 2025. Elaboración propia basada en el recuento total de resultados obtenidos bajo la búsqueda "*artificial intelligence COPD*".

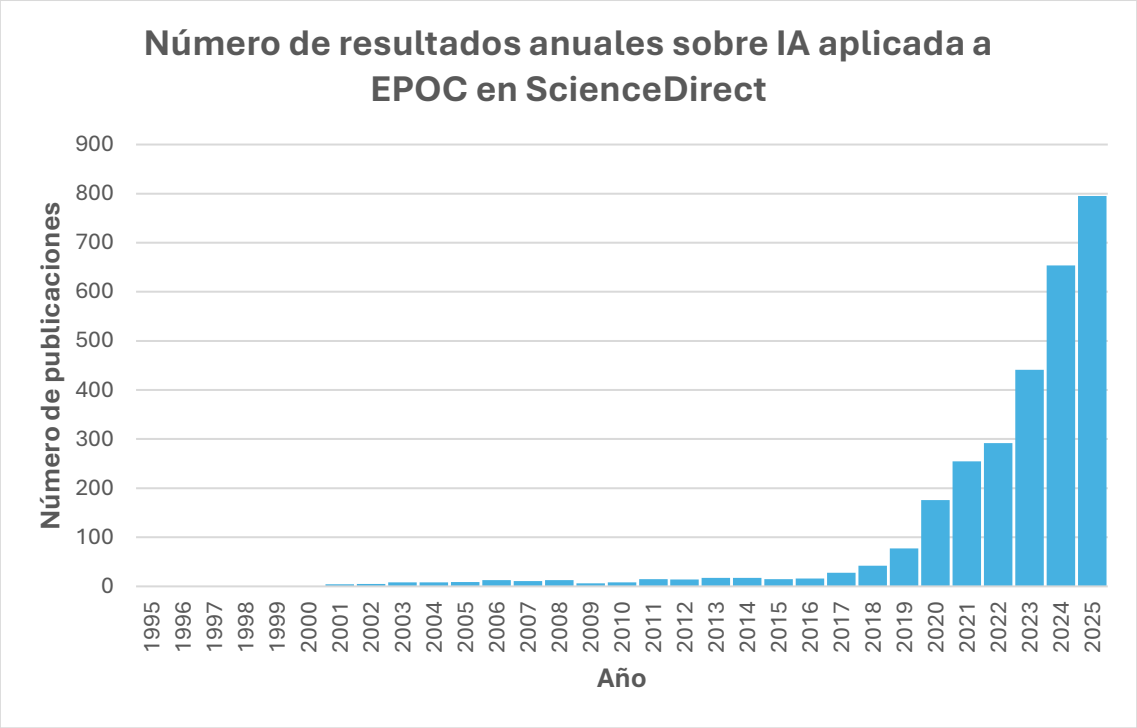

Figura 14. Número de publicaciones en *ScienceDirect* sobre Inteligencia Artificial en EPOC entre 1995 y 2025. Elaboración propia basada en el recuento total de resultados obtenidos bajo la búsqueda *"artificial intelligence COPD"*.

En la Figura 15 se presentan distintas aplicaciones de la IA para el abordaje de la EPOC, observando su integración en diferentes aspectos de la enfermedad [42].

La IA se alimenta de datos procedentes de exploraciones físicas, que brindan información a través de cuestionarios, evaluaciones médicas y pruebas como la caminata de seis minutos (6MWT, *Six Minutes Walking Test*). Esta información se complementa con variables fisiológicas obtenidas por medio de distintos dispositivos o pruebas, como la frecuencia respiratoria (RR, *respiratory rate*), presión sanguínea (BP, *blood pressure*), frecuencia cardíaca (HR, heart rate), sonidos pulmonares (LS, *lung sounds*), además de resultados hematológicos y de tomografía computarizada (TC), especialmente de tórax. También es importante tener en cuenta factores como la genética, las condiciones meteorológicas y el ambiente, que pueden influir significativamente en el desarrollo de la enfermedad.

La combinación de toda esta información permite aplicar modelos de predicción de exacerbaciones agudas (AECOPD) mediante *Machine Learning*. Además, el uso conjunto de *Machine Learning* y *Deep Learning* proporciona asistencia en el diagnóstico, lo que da lugar a otra aplicación de la IA en esta enfermedad: el diagnóstico inteligente.

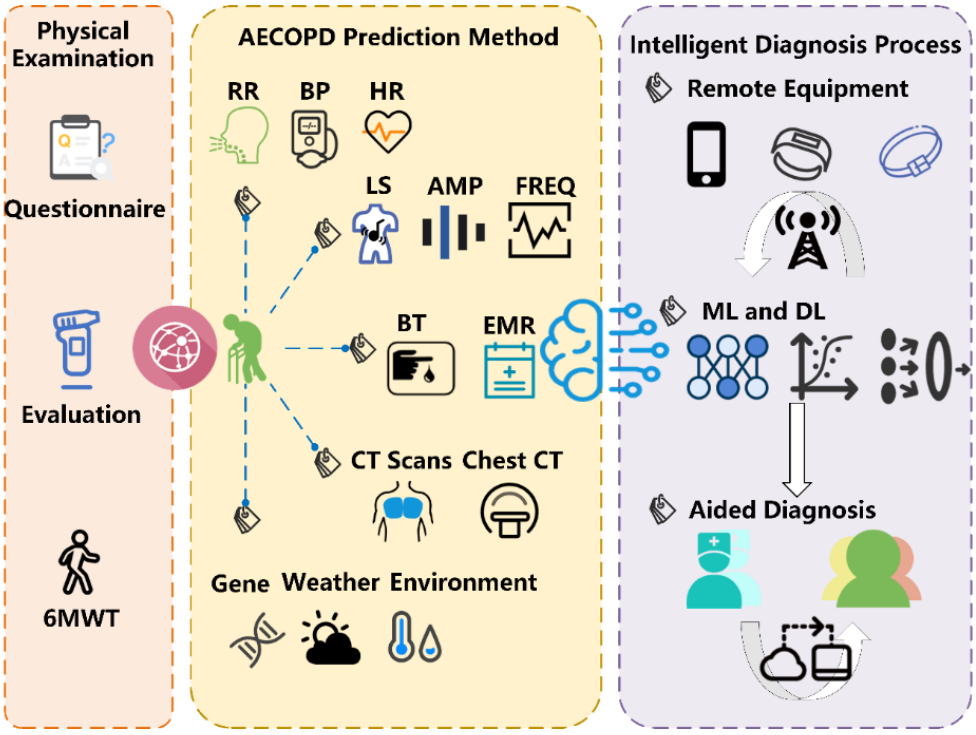

Figura 15. Aplicaciones de la Inteligencia Artificial en EPOC [42].

6MWT:6 Minutes Walking Test. RR: respiratory rate. BP: blood pressure. HR: heart rate. LS: lung sound. AMP: amplitude. FREQ: frequency. BT: blood test. EMR: electronic health records.

Sin embargo, más allá del diagnóstico, uno de los desafíos más relevantes en el abordaje de la EPOC es el manejo de los reingresos por exacerbaciones. Estos eventos conllevan un deterioro en la calidad de vida del paciente, un aumento sustancial de los costes sanitarios, mayor riesgo de complicaciones derivadas de la enfermedad e incremento de la tasa de mortalidad en comparación con la población general. Por consiguiente, resulta clave identificar los factores de riesgo que predisponen a estas agudizaciones y elaborar modelos predictivos simples, eficaces y aplicables durante la práctica clínica. Dichos modelos son herramientas con un gran potencial, ya que permiten anticipar el posible deterioro de la enfermedad. Con ello, se favorecen intervenciones más tempranas y tratamientos adaptados al perfil clínico del paciente, contribuyendo significativamente a prolongar su supervivencia y evitar dificultades en su vida diaria [43].

No obstante, la EPOC es una enfermedad muy heterogénea y actualmente no está lo suficientemente reconocida. Esto complica la detección temprana y compromete la fiabilidad de las predicciones. A pesar de que se han desarrollado modelos predictivos de reingresos por exacerbación con una amplia variedad de características, sus rendimientos y validez clínica continúan siendo limitados [43].

Por ello, el presente trabajo pretende proporcionar una nueva aproximación mediante un modelo predictivo de reingresos por exacerbaciones en EPOC, aplicando técnicas de IA sobre

variables clínicas recogidas en el ámbito hospitalario. Así, se pretende contribuir a mejorar el abordaje de un problema de gran impacto social y económico cuya solución podría marcar una gran diferencia en el pronóstico de los pacientes.

CAPÍTULO 2. HIPÓTESIS Y OBJETIVOS

2.1. Hipótesis

El presente trabajo se sustenta en:

(1) Hipótesis clínica:

Se espera que las variables de estudio recopiladas en el ámbito hospitalario contengan información relevante asociada al reingreso de pacientes por exacerbación de EPOC en los 30 días después del alta.

(2) Hipótesis técnica:

Desde el punto de vista técnico, se espera que los modelos de aprendizaje computacional entrenados con un subconjunto óptimo de las variables de estudio sean capaces de predecir los reingresos hospitalarios de pacientes de EPOC.

2.2. Objetivos

El **objetivo principal** consiste en desarrollar y validar un modelo predictivo de reingreso hospitalario por exacerbación en pacientes con EPOC en un periodo de 30 días posterior al alta, aplicando técnicas computacionales de aprendizaje automático.

Con el fin de alcanzar este objetivo general, se detallan a continuación los siguientes **objetivos específicos**:

I. Determinar el subconjunto de variables más relevantes asociadas con el reingreso temprano (30 días) tras el alta hospitalaria en pacientes ingresado por exacerbación de EPOC.

II. Desarrollar un modelo predictivo de reingreso mediante el enfoque de *ensamble learning Random Forest* (RF) que tome a su entrada las variables óptimas seleccionadas.

III. Comparar el rendimiento predictivo del modelo RF con el alcanzado por una red neuronal perceptrón multicapa (MLP), que se tomará como *benchmark* de referencia.

IV. Validar de forma independiente (validación temporal) los modelos diseñados en una nueva población de estudio recopilada prospectivamente, pudiendo valorar su capacidad de generalización.

CAPÍTULO 3. SUJETOS Y VARIABLES DE ESTUDIO

3.1. Aspectos éticos

Este Trabajo de Fin de Grado (TFG) ha sido elaborado cumpliendo rigurosamente con los principios éticos y legales pertinentes. Se emitió un dictamen favorable por parte del Comité de Ética de la Investigación con medicamentos (CEIm) procedente de las Áreas de Salud de Valladolid (Ref.: PI-24-672-H, de 18 de diciembre de 2024) que se rige según las normas de Buenas Prácticas Clínicas (BPC, CMP/ICH/135/95).

Asimismo, las bases de datos recopiladas fueron tratadas conforme al Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo relativo a la protección de personas físicas en cuanto al tratamiento de sus datos personales y la libre circulación de estos. También se actuó de acuerdo con la Ley Orgánica 3/2018 sobre Protección de Datos Personales y garantía de los derechos digitales (LOPDPGDD). Por ello, se mantuvo absoluta confidencialidad a lo largo de todo el estudio.

Además, todos los pacientes firmaron un consentimiento informado, a través del cual se les explicó en detalle la finalidad del estudio y su derecho a abandonarlo en cualquier momento sin consecuencia alguna.

3.2. Diseño del estudio

El presente trabajo es un estudio observacional ambispectivo de diseño y validación de modelos predictivos, con una fase retrospectiva (datos recopilados entre octubre de 2017 y junio de 2019) y una etapa prospectiva (datos recopilados entre enero y junio de 2025). Ambos *datasets* contienen información procedente del Servicio de Neumología del Hospital Universitario Río Hortega de Valladolid. Los pacientes incluidos debían cumplir los siguientes **criterios de inclusión**:

- Diagnóstico previo de EPOC confirmado por pruebas clínicas.
- Ingreso hospitalario por exacerbación de la enfermedad.
- Edad superior a 18 años.

Como **criterios de exclusión**, se establecieron únicamente la ausencia de consentimiento informado y no haber cumplido la mayoría de edad.

3.3. Tamaños muestrales

En cuanto a la **base de datos retrospectiva**, se calculó un tamaño muestral mínimo de 176 pacientes. Esta estimación se efectuó mediante el *software G*Power 3*.1 (Heinrich Heine Universität Düsseldorf, Germany) para una potencia estadística del 90% y un tamaño de efecto de 0.405. Finalmente, se reclutaron 246 pacientes consecutivos, de los cuales 204 no reingresaron en el periodo de seguimiento y 42 reingresaron dentro de los 30 días posteriores al alta.

Respecto a la **base de datos prospectiva**, el tamaño muestral estimado fue calculado como un porcentaje del estudio original en base a las particiones comúnmente empleadas en el diseño de modelos predictivos (70% de instancias para el entrenamiento y 30% para *test*). Finalmente, para este conjunto de validación se reclutaron un total de 75 pacientes consecutivos, de los cuales 70 no reingresaron y 5 fueron rehospitalizados dentro de los 30 días de seguimiento.

3.4. Variables de estudio

Las bases de datos recogidas se construyeron a partir de la información registrada en la historia clínica de los pacientes, incluyendo sus antecedentes generales y los relativos a su EPOC. Asimismo, se integraron los datos generados en la práctica clínica habitual durante el ingreso por exacerbación.

Los pacientes fueron caracterizados de forma global a través de un total de 229 variables clínicas, incluyendo fechas, la caracterización del reingreso o del fallecimiento, si estos se producían. La variable dependiente o *target* codificó el evento "reingreso", tomando valores binarios (0/1) atendiendo a si se produce (1) o no (0) un reingreso por exacerbación de EPOC durante los 30 días posteriores al alta.

A partir de las 229 variables de partida, se identificó un conjunto inicial de variables predictoras en base al conocimiento de expertos y hallazgos del estado-del-arte (*knowledge-based*) compuesto por 158 características, agrupadas en los siguientes campos:

- **Datos sociodemográficos y antropométricos (12):** edad, sexo, procedencia, estado civil, hogar, estudios, actividad laboral, movilidad y tipo de cuidador, peso, altura e IMC (índice de masa corporal).

- **Hábitos (3):** tabaquismo, índice tabáquico y alcoholismo.

- **Datos clínicos (5):** grupo de riesgo clínico (GRC), anticoagulación, antiagregación, vacuna antigripal (año previo) y vacuna antineumocócica.

- **Comorbilidades (18):** hipertensión arterial, diabetes mellitus, dislipemia, cardiopatía isquémica, insuficiencia cardíaca, insuficiencia respiratoria, bronquiectasias, taquiarritmia, accidente cerebrovascular, neoplasia (pulmón, otras), enfermedad renal, osteoporosis, ansiedad, depresión, anemia, tromboembolismo pulmonar y síndrome de apnea-hipopnea.

- **Espirometría previa al ingreso (9):** FVC basal (capacidad vital forzada), FEV1 basal (volumen espiratorio forzado en el primer segundo), FEV1/FVC basal. De cada uno de los tres parámetros se obtuvo su valor absoluto, porcentaje sobre el valor teórico y normalización z-score.

- **Caracterización de la gravedad de la enfermedad (5):** estratificación del riesgo, fenotipo clínico según GesEPOC, grado de obstrucción al flujo aéreo, estadio según GOLD y número de exacerbaciones por agudización de EPOC en el año previo.

- **Terapia basal (20):** oxigenoterapia continua domiciliaria, ventilación no invasiva, medicación inhaladora (ninguna, SABA, SAMA, LABA, LAMA y corticoides), medicación de rescate (ninguna, SABA, SAMA, LABA, LAMA y corticoides), corticoterapia sistémica continua, teofilinas, IFDE4 (inhibidores de la fosfodiesterasa tipo 4), mucolíticos, antibióticos y rehabilitación respiratoria.

- **Test de evaluación de influencia de la EPOC en diversos ámbitos de la vida diaria (13):**
  - **Cuestionario de Barthel (1):** puntuación total. Este permite analizar el nivel de dependencia [44].
  - **Cuestionario CAT (1):** puntuación total. Mide el impacto en la vida diaria (descrito previamente).
  - **Cuestionario EuroQOL-5D (5):** puntuación del grado de movilidad, cuidado personal, dificultades en actividades cotidianas, dolor o malestar, y ansiedad o depresión. Por tanto, este test proporciona una medición de la calidad de vida relacionada con la salud (CVRS) [45].
  - **Cuestionario TAI (4):** puntuación del nivel de adhesión e incumplimientos errático, deliberado e inconsciente. Se trata del único test específico para medir la adhesión a los inhaladores [46].
  - **Índice de Comorbilidad de Charlson (1):** puntuación total. Este evalúa la supervivencia a los diez años [47], atendiendo a la edad del sujeto y sus comorbilidades.
  - **Escala mMRC (1):** puntuación total. Mide el grado de disnea (explicado con anterioridad).

- **Duración y motivo de ingreso (5):** número de días ingresado, causa de ingreso (infecciosa, infecciosa debido a bacterias, infecciosa debido a virus y presencia de microorganismos resistentes).

- **Estado del paciente al ingreso (12):**
  - **Constantes vitales (3):** tensión sistólica, diastólica y frecuencia cardíaca.
  - **Gasometría (4):** pH, PCO_2 (presión parcial de dióxido de carbono en la sangre), PO_2 (presión parcial de oxígeno en la sangre) y HCO_3 (concentración de iones de bicarbonato).
  - **Analítica (5):** leucocitos, neutrófilos (valor absoluto y porcentaje) y eosinófilos (valor absoluto y porcentaje). El porcentaje hace referencia a la proporción relativa respecto al número de leucocitos totales.

- **Síntomas y complicaciones al ingreso (21):**
  - **Síntomas clínicos (7):** aumento de tos, disnea, expectoración, purulencia del esputo, dolor torácico, fiebre y días de clínica que motivan el ingreso.
  - **Signos (7):** uso de musculatura accesoria, movimientos torácicos paradójicos, cianosis, edemas periféricos, inestabilidad hemodinámica, deterioro del estado mental y disnea.
  - **Complicaciones asociadas (7):** arritmias, insuficiencia cardíaca, cardiopatía isquémica, derrame pleural, neumonía, sepsis e insuficiencia respiratoria.

- **Terapia al ingreso (15):** medicación inhaladora (ninguna, SABA, SAMA, LABA, LAMA y corticoides), corticoterapia sistémica, ingreso en UVI, oxigenoterapia, pauta de teofilinas, IFDE4, mucolíticos, ventilación (no invasiva e invasiva) y antibioterapia.

- **Terapia al alta (20):** oxigenoterapia continua domiciliaria, ventilación no invasiva, medicación inhaladora (ninguna, SABA, SAMA, LABA, LAMA, corticoides), medicación de rescate (ninguna, SABA, SAMA, LABA, LAMA, corticoides), corticoterapia sistémica continua, uso de antibióticos, teofilinas, IFDE4, mucolíticos y rehabilitación respiratoria.

CAPÍTULO 4. METODOLOGÍA

4.1. Esquema general de trabajo

Como se ha expuesto con anterioridad, para la elaboración de este estudio se cuenta con dos bases de datos diferentes: una retrospectiva y otra prospectiva. A continuación, en este apartado, se expone el flujo general de trabajo adoptado.

Partiendo de la base de datos retrospectiva, en primer lugar, se efectuó el tratamiento de datos o *data curation.* Posteriormente, mediante la aplicación del algoritmo ReliefF que se explicará detalladamente en apartados posteriores, se seleccionaron las variables predictoras de mayor relevancia, que constituyen las entradas comunes tanto para los modelos predictivos a desarrollar como para la futura validación temporal mediante la base de datos prospectiva. De este modo, en el procedimiento de recopilación prospectivo fue posible centrar los esfuerzos únicamente en las variables más relevantes, proporcionando agilidad en el proceso de recolección, transcripción y construcción de la base de datos prospectiva. Cabe destacar que la generación de esta cohorte se realizó de forma paralela al diseño y desarrollo de los modelos predictivos.

Una vez determinadas las características de mayor peso en el estudio, se procedió a diseñar y optimizar los modelos sobre la base de datos retrospectiva: *Random Forest* y la red MLP. Estos fueron evaluados mediante el cálculo de diversas métricas de rendimiento que se especificarán en este capítulo. Tras esto, los modelos se validaron aplicándolos sobre la base de datos prospectiva, sin efectuar una nueva optimización de los hiperparámetros. En su lugar, se aplicaron los mismos valores que los establecidos previamente con el conjunto retrospectivo y su rendimiento fue caracterizado atendiendo a las mismas métricas que las obtenidas en la base de datos retrospectiva. Esta validación temporal permitió conocer la capacidad de generalización de los modelos sobre un conjunto de datos diferente independiente.

La Figura 16 resume en forma de diagrama el flujo de trabajo descrito.

4.2. *Data curation*

4.2.1. Exploración de datos perdidos

El punto de partida de este trabajo es la base de datos retrospectiva, que contiene la información de las variables previamente descritas y que fue transcrita en Excel para su análisis y tratamiento posterior. Sin embargo, se identificaron valores faltantes, definidos como aquellos que no han sido almacenados para una variable en la muestra de interés. Su correcta gestión resulta

fundamental para evitar los múltiples problemas derivados de los mismos, que pueden condicionar significativamente los resultados de cualquier estudio [48].

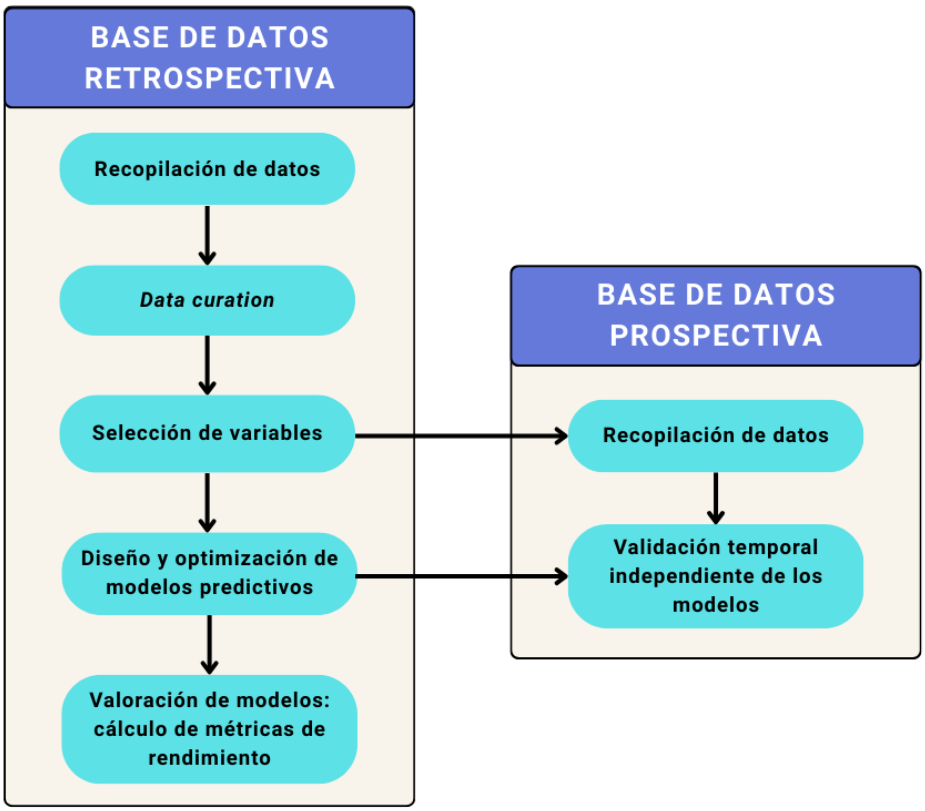

Figura 16. Diagrama de flujo de trabajo adoptado en el estudio.

Entre sus efectos negativos se incluyen la reducción de la potencia estadística, posibles sesgos en las estimaciones de parámetros, menor representatividad de las muestras y dificultades en el análisis. Estos factores comprometen la validez, conclusiones y robustez del trabajo, subrayándose la importancia de un tratamiento adecuado [48].

En este trabajo, se aplicó un proceso estructurado para garantizar la correcta gestión de los datos faltantes durante el preprocesamiento de la base de datos retrospectiva. A continuación, se describen las estrategias implementadas:

- **Eliminación de pacientes fallecidos**. En primer lugar, se excluyeron de los datos aquellos pacientes cuyo motivo de alta fue el fallecimiento durante el ingreso. Dado que se pretende predecir los reingresos durante los 30 días posteriores al alta, no se ha podido efectuar un seguimiento de los mismos durante dicho periodo.

- **Obtención inicial de valores ausentes.** Tras prescindir de las filas que representan pacientes fallecidos, se procedió a calcular el porcentaje de datos perdidos o *NaNs (Not a Number)* tanto por cada paciente (fila), como por cada columna (variable recogida). Esta etapa facilitó una visión preliminar de la distribución y alcance de los datos faltantes.

- **Exclusión de columnas con elevado porcentaje de *NaNs* y baja varianza.** Una vez conocidas visualmente aquellas variables cuyo porcentaje de datos ausentes era superior al umbral establecido, se procedió a excluirlas. Asimismo, aquellas cuyo sumatorio total de muestras fuese 0 o 1 también fueron eliminadas del estudio. Esta decisión se basa en que su variabilidad es insuficiente para contribuir a la discriminación entre la clase positiva y negativa y, por consiguiente, su inclusión podría afectar la capacidad predictiva del modelo.

- **Actualización del porcentaje de datos perdidos por paciente.** Tras el filtrado previo de columnas, se reevaluaron los datos ausentes por cada paciente o fila, ya que, al haberse eliminado variables, las estimaciones previas de pérdida de datos dejarían de tener validez. En este caso, se optó por no establecer un umbral con el que eliminar filas, puesto que el porcentaje de datos perdidos resultó ser bajo. Además, dado el limitado número de pacientes del estudio, se priorizó la preservación de la mayor cantidad posible de registros. De este modo, se maximiza la representatividad de los datos.

4.2.2. Imputación de datos: *K* vecinos más cercanos (KNN)

El manejo de datos perdidos puede abordarse mediante distintas estrategias, como la eliminación de instancias o el reemplazo por valores estimados. Esta última actuación se conoce como imputación, para la cual se dispone de diversas técnicas [49].

En este trabajo, se aplicó el algoritmo *K* vecinos más cercanos *(K-Nearest Neighbors,* KNN*),* un método no paramétrico de aprendizaje supervisado creado en 1951 por los estadísticos Evelyn Fix y Joseph Hodge [50]. Esta herramienta identifica a los *K* vecinos más cercanos de los valores ausentes y emplea estos para poder realizar la imputación atendiendo a una distancia entre las instancias [49].

A pesar de que existen diferentes métricas de distancia (Manhattan, Mahalanobis o coeficientes de correlación) [50], se optó por aplicar la distancia euclídea en este caso, ya que ha demostrado mayor eficiencia y resulta ser la más habitual en la literatura [49].

Dado que se tienen filas o pacientes con datos incompletos, esta métrica se implementó únicamente con las variables no vacías de estos casos. Además, se prescindió de la variable objetivo (reingreso), puesto que no debe participar en la etapa de imputación.

Ante el carácter mixto de las variables contenidas en esta base de datos retrospectiva, fue preciso discriminar entre variables continuas y categóricas. Así, para las primeras citadas se calculó la media de los *K* vecinos más cercanos (*mean rule*), mientras que para las segundas se obtuvo la moda *(majority rule)* [51].

Por otra parte, cada valor ausente a estimar se imputó atendiendo únicamente a aquellos vecinos pertenecientes a la misma clase (reingreso o no reingreso). El fin de esta actuación es prevenir mezclas entre clases que puedan incorporar valores no representativos del grupo de pacientes del que forma parte dicho sujeto con datos incompletos.

Respecto al número de vecinos, se asignó un valor de *K = 5,* decisión respaldada por su eficacia en otros estudios previos, donde también se aplicó KNN [52-55]. El uso de *K > 1* contribuye a controlar el ruido en los datos, mientras que valores demasiado bajos podrían introducir una gran sensibilidad a valores atípicos. No obstante, un número excesivamente elevado provoca una distorsión en la distribución de los datos [55]. Dado que se ha apreciado que los valores *K = 3, 5, y 7* ofrecen diferencias mínimas en términos de rendimiento [54], y teniendo en cuenta las consideraciones previas, se decidió asignar un valor intermedio razonable y ampliamente contrastado (*K= 5*).

4.3. Análisis descriptivo

Tras el tratamiento de los datos perdidos y la imputación de estos, se procedió a efectuar un análisis descriptivo, cuyo objetivo es resumir las variables mediante métricas representativas y visualizar los datos. De este modo, es posible identificar tendencias, patrones y alcanzar una mayor comprensión de la información disponible [56]. Para este análisis, se calcularon los estadísticos descriptivos convencionales de tendencia central y dispersión, adoptando distintas estrategias en función de la naturaleza de la variable, es decir, si es continua o categórica.

Respecto a las continuas, se calculó la mediana, primer cuartil (Q1), tercer cuartil (Q3) y rango intercuartílico (IQR).

- **Mediana**: es el valor que divide un conjunto de datos ordenados de menor a mayor en dos subconjuntos iguales. Es decir, la mitad de las observaciones se encuentran por debajo de la mediana y el otro 50% por encima. Además, dependiendo del tamaño muestral [57]:
    - Si el número de datos (*n*) es impar, la mediana coincide con el dato central.
    - Si *n* es par, la mediana se obtiene como la media de los dos datos centrales.

- **Cuartiles (Q1, Q2, Q3)**: son medidas de posición de tendencia no central que dividen la población de estudio en cuatro partes iguales. Q1 se corresponde con el valor por debajo

del cual se sitúa el 25% de las observaciones; Q2 (también denominado mediana) deja el 50% de los datos por debajo de él; y Q3, el 75% [57].

- **Rango intercuartílico (IQR):** medida de dispersión absoluta que representa la distancia entre el primer y tercer cuartil (*Q3-Q1).* Es un estadístico robusto, poco sensible a valores atípicos y muy útil para observar la variabilidad central de las muestras [58], [59].

Para el análisis de las variables categóricas, se han obtenido las frecuencias absolutas (número total de observaciones por categoría) y relativas (porcentajes).

4.4. Selección de variables

Tras conseguir una base de datos con todas las instancias completas, se procedió a seleccionar las variables que constituirán las entradas de los modelos predictivos a desarrollar. Uno de los desafíos más importantes en minería de datos es la caracterización de las relaciones entre una o más variables y la variable objetivo (reingreso en este trabajo).

Aunque se tengan conjuntos de datos comprendidos por varias características, en la mayoría de los casos tan solo una proporción de las mismas resulta relevante. El problema es que es muy difícil conocer *a priori* estas variables irrelevantes y su inclusión provoca un incremento de la complejidad y de la carga computacional. Por ello, es necesaria la aplicación de métodos que permitan identificar las características de mayor peso en un estudio y descarten aquellas que podrían dificultar posteriores tratamientos y no contribuyan a aportar información valiosa adicional.

Con este fin, se ha implementado el algoritmo *ReliefF*, que permite calcular una estadística para cada variable que refleje su calidad o relevancia respecto a la variable objetivo reingreso. Estas estadísticas se denominan pesos o puntuaciones y pueden adquirir valores dentro del intervalo [-1, 1], siendo -1 la relevancia mínima y 1 la máxima [60], [61].

La aplicación de *ReliefF* precisa la asignación de un número de vecinos *(K).* Para cada instancia del *dataset*, se efectúa una búsqueda de los *K* vecinos más cercanos que sean similares y pertenezcan a la misma categoría (denominados *hits*), así como de los *K* vecinos más cercanos que también sean similares, pero pertenezcan a una clase diferente (llamados *misses*) [60].

El algoritmo compara las variables entre la instancia o fila actual y las correspondientes a sus *K* vecinos más cercanos. Si su valor difiere de uno de los *misses,* esto refleja que dicha característica contribuye a discriminar entre clases, por lo que su peso o puntuación aumenta. Por el contrario, si el valor de la instancia actual difiere respecto al de uno de los *hits*, dicha variable no permite distinguir entre categorías y el peso de la misma disminuye.

El peso calculado para una variable *A* (*W[A]*) puede definirse desde un punto de vista probabilístico, siendo este la diferencia entre dos probabilidades condicionadas: la probabilidad de que un vecino de tipo *miss* posea un valor distinto de *A* y la probabilidad de que un vecino de clase *hit* tenga un valor diferente [60].

$$W[A] = P\ (valor\ distinto\ de\ A \mid vecino\ miss) - P\ (valor\ distinto\ de\ A \mid vecino\ hit) \quad (4.1)$$

Las variables predictoras del conjunto de datos fueron previamente estandarizadas para evitar escalas diferentes.

Atendiendo a las mismas recomendaciones detalladas para determinar el valor más adecuado de *K* en la etapa de imputación de datos (es decir, evitar tanto valores muy pequeños, que podrían ser afectados por espúreos, como valores muy grandes, que podrían diluir las diferencias), se fijó *K = 5.*

Finalmente, la selección de características se efectuó sobre la lista ordenada de variables según la puntuación otorgada por *ReliefF*. Siguiendo la regla empírica de 10 instancias por variable y dado que se disponía de una muestra de 243 pacientes, se seleccionaron las 24 variables con mayor peso, descartando el resto.

4.5. Desarrollo de modelos predictivos

Una vez seleccionadas las características que guardan mayor relación con la variable objetivo (reingreso), se desarrollaron los modelos predictivos capaces de predecir los reingresos por exacerbación de EPOC en los 30 días posteriores al alta. En este TFG se confeccionaron dos enfoques de *Machine Learning:* un modelo principal basado en *Random Forest* y una red perceptrón multicapa (MLP).

Para el diseño y optimización de los modelos se utilizó la base de datos retrospectiva previamente depurada, conformada únicamente por las 24 variables predictoras seleccionadas y un total de 243 pacientes. Este conjunto de partida se dividió en un grupo de entrenamiento (*train1*), que incluye el 70% de los pacientes; y un grupo de prueba (*test*), que se corresponde con el 30% restante. A su vez, el conjunto de entrenamiento se dividió en un grupo de entrenamiento final (*train2*), que representa el 70% del mismo; y un conjunto de validación conformado por el 30% restante para así optimizar los hiperparámetros pertinentes. A través de este entrenamiento y validación interna, se calcularon finalmente métricas que permitieron caracterizar el rendimiento de los modelos y comparar su capacidad predictiva. En la Figura 17, se ilustra de manera esquemática estas particiones y el número de pacientes resultante en cada uno de los conjuntos:

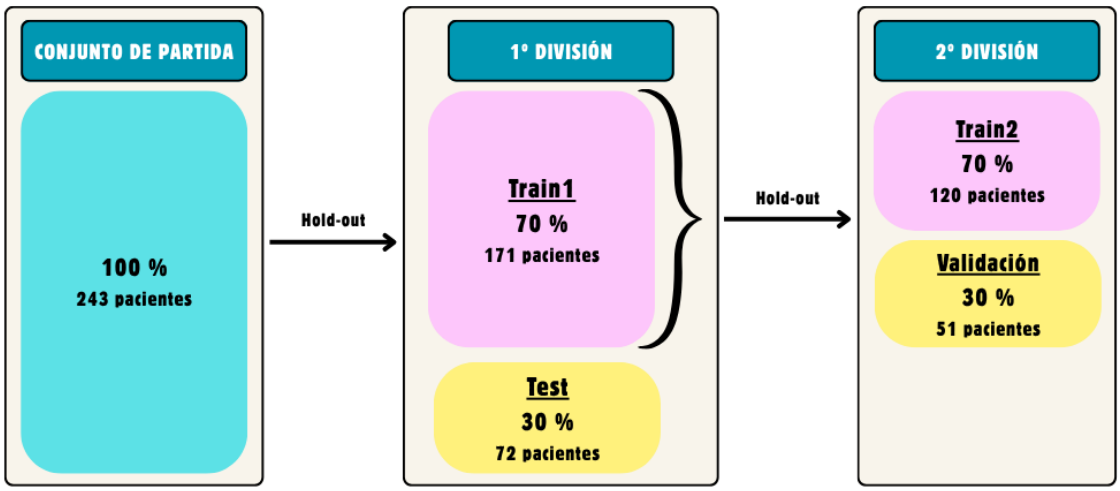

Figura 17. Divisiones de la población bajo estudio a lo largo del desarrollo de los modelos predictivos.

4.5.1. *Random Forest*

Random Forest (RF) es, como se ha adelantado previamente, la técnica de *Machine Learning* aplicada para desarrollar el modelo predictivo principal de este TFG. Esta herramienta fue elaborada por Breiman en el año 2001 y se basa en la generación de múltiples árboles de decisión aleatorios sobre un conjunto de datos de entrenamiento [62]. Este algoritmo forma parte de los *ensemble methods*, es decir, métodos que combinan las predicciones de diferentes estimadores proporcionando resultados más robustos y una mejora en la capacidad de generalización del modelo [63].

En este caso en particular, RF reduce la varianza y evita el sobreajuste en comparación con los árboles de decisión individuales [63], [64]. Cada árbol contiene un subconjunto de variables predictoras *m* que debe ser inferior al número total de características predictoras *M* (es decir, debe cumplirse la condición $m < M$) [62]. Asimismo, mediante la aplicación de *Bootstrap*, cada árbol es entrenado a partir de instancias (pacientes en este estudio) seleccionadas aleatoriamente.

Cabe destacar que, las variables e instancias atribuidas a uno de los árboles generados pueden ser comunes a los asignados al resto. Es decir, pueden repetirse características y pacientes entre diferentes árboles de decisión. Sin embargo, existe la posibilidad de que algunas de las instancias no hayan sido asignadas a ninguno de estos. Estas son denominadas popularmente como *out of the bag* (OBB) y se utilizan para validaciones internas automáticas [62].

Finalmente, las salidas de todos los árboles creados se combinan para proporcionar la salida final del modelo, conocida como ensamblado. Para el cálculo de la misma, debe realizarse una distinción sobre la variable objetivo a estimar. Si esta es continua, RF estaría aplicándose a un problema de regresión y la salida final se obtendría normalmente mediante el promedio de las

predicciones de todos los árboles. Por el contrario, si la variable *target* es categórica (reingreso/no reingreso), se trataría de un problema de clasificación y se aplicaría el voto por mayoría [62].

A continuación, en la Figura 18 se ilustra gráficamente el funcionamiento del algoritmo *Random Forest* [62]:

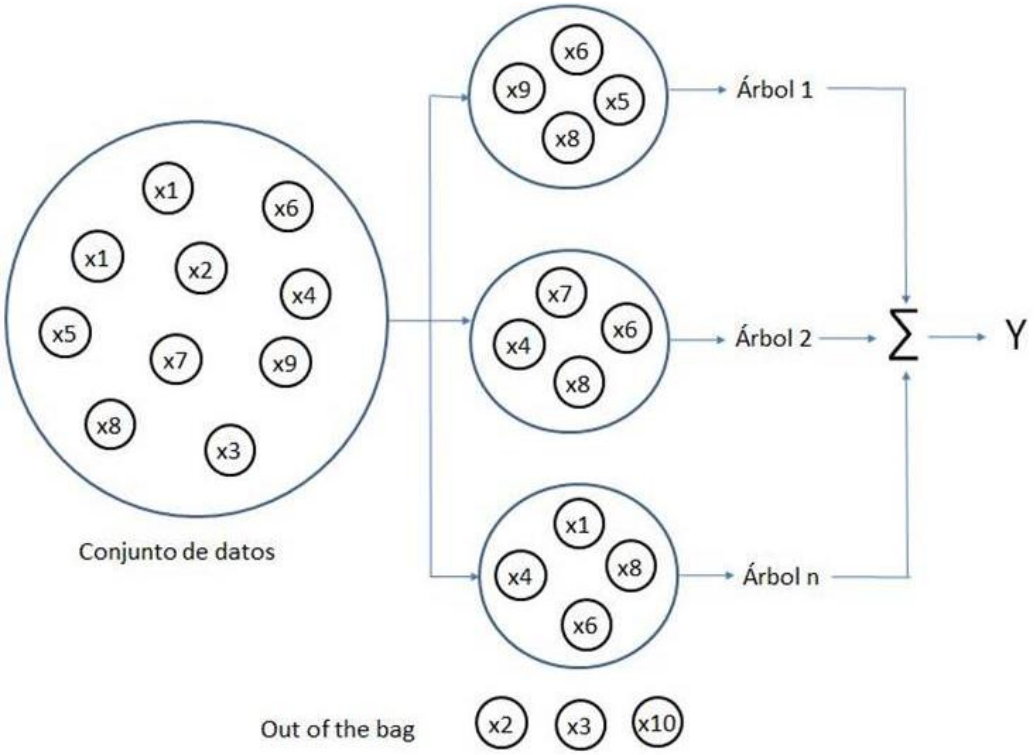

Figura 18. Esquema del funcionamiento de *Random Forest* [62].

En este esquema se resume visualmente cómo RF selecciona aleatoriamente diferentes subgrupos de pacientes procedentes del conjunto de datos inicial y se aplican sobre los árboles de decisión confeccionados, devolviendo finalmente una salida conjunta *Y*. En la parte inferior se incluyen tres instancias que, como puede apreciarse, no han sido asignadas a ningún estimador (*out of the bags*).

Random Forest es actualmente el método de *Machine Learning* más implementado en modelos predictivos. Este hecho es gracias a las múltiples ventajas que proporciona y que respaldan su idoneidad [64], [62]:

- Adaptación natural tanto a problemas de regresión como de clasificación.
- Velocidad óptima de entrenamiento y predicción.
- Desempeño óptimo sobre bases de datos de elevada dimensionalidad.
- Mayor simplicidad para ser entrenado respecto a herramientas de mayor complejidad, pero con un rendimiento muy similar.

Estas ventajas motivaron la inclusión de esta técnica en el presente estudio. Para su implementación, se dispone de diversos hiperparámetros configurables por el usuario, que consisten en variables definidas de antemano que permiten optimizar el entrenamiento del modelo de aprendizaje automático. Su adecuado ajuste garantiza un buen desempeño y una gran eficiencia computacional. En consecuencia, se llevó a cabo la optimización de los siguientes hiperparámetros:

- **Número de árboles de decisión.** A mayor cantidad de estos estimadores, mejor rendimiento, pero mayor carga computacional. Además, a partir de un número de árboles, el modelo habrá alcanzado su rendimiento máximo, por lo que dejará de mejorar [63].

- **Coste.** Generalmente, se trata de una matriz cuadrada en la que las filas se corresponden con las clases verdaderas (*true class*), mientras que las columnas son las clases predichas (*predicted class)* [65]. El objetivo de esta matriz de costes es definir las penalizaciones asociadas a los errores en la clasificación [63]. Dada la naturaleza desbalanceada de los datos disponibles, la configuración de este hiperparámetro permitió asignar un coste o penalización mayor a la predicción errónea de "no reingreso" cuando la clase verdadera es "reingreso". De este modo, se otorga más peso a la clase positiva, ya que esta es la minoritaria. Así, se consigue mitigar los efectos de dicho desbalance tan marcado (aunque común en el ámbito clínico), que puede afectar negativamente a la sensibilidad del modelo. La estructura general de esta matriz de costes se muestra en la Tabla 9:

Tabla 9. Estructura de la matriz de costes para *Random Forest.*

CLASE DE DATOS		**Clase predicha (*predicted class*)**	
		0 = No reingreso	**1 = Reingreso**
Clase verdadera (*true class*)	**0 = No reingreso**	0	Coste de FP
	1 = Reingreso	Coste de FN	0

FP: falsos positivos; FN: falsos negativos.

Como puede observarse, la diagonal principal de la matriz presenta valores nulos, puesto que, en esos casos, la clase predicha coincide con la verdadera. Por consiguiente, el modelo no habría cometido ningún error y no tendría sentido una penalización distinta de cero.

- **Tamaño mínimo de hoja.** Número mínimo de observaciones por cada hoja [65].

- **Número máximo de divisiones.** Está relacionado directamente con la profundidad máxima de los árboles. Dado que los modelos cuyos estimadores son árboles tienen mayor riesgo de sobreajuste, no es recomendable profundidades grandes [63].

- **Número de predictores.** Con este hiperparámetro, se determina el número de características predictoras elegidas aleatoriamente para cada división en los árboles de decisión [65]. Para cada nodo del árbol que se está generando, se selecciona con probabilidad uniforme un subconjunto de variables en base al conjunto inicial. Entre las características seleccionadas, el algoritmo aplica, atendiendo a un criterio estadístico como el índice de Gini o la entropía, la división que produce la mejor partición de los datos. A menor valor de este hiperparámetro, mayor será la reducción de la varianza, pero el sesgo se verá aumentado [63].

Respecto a la dinámica de optimización, esta se efectuó de manera secuencial. Para cada uno de los hiperparámetros previamente mencionados, se estableció un rango de valores predefinido y se entrenó el modelo con *train2,* probando cada valor de dicho intervalo mientras los demás hiperparámetros se mantuvieron fijos en valores acordes con el contexto y lo observado en la literatura. En la primera optimización (número de árboles), los valores fijados fueron arbitrarios, mientras que, en las sucesivas iteraciones, se aplicaban los hiperparámetros optimizados hasta el momento de forma secuencial. La métrica *F1 score* fue la empleada para guiar la búsqueda del valor óptimo, calculada a partir de la matriz de confusión del conjunto de validación. Además de ajustar los hiperparámetros, se optimizó el umbral de decisión en la predicción atendiendo al mismo criterio de maximización de *F1 score,* pero aplicando la curva ROC para conocer el rendimiento en distintos umbrales.

Tras identificar los valores óptimos, se reentrenó el modelo final con el conjunto de entrenamiento *train1* para posteriormente ser evaluado en el conjunto *test*.

4.5.2. Red neuronal perceptrón multicapa *(Multi-Layer Perceptron*, MLP)

En 1969, Minsky y Papert manifestaron las dificultades que presentaba el perceptrón simple para la resolución de problemas no lineales. No obstante, se observó que la combinación de varios perceptrones podía cubrir sus limitaciones [66].

Rumelhart y sus colaboradores presentaron en 1986 la Regla Delta Generalizada, un algoritmo para entrenar redes neuronales. A partir del mismo fue posible adaptar los pesos mediante retropropagación del error y se extendió a arquitecturas con múltiples capas y funciones de activación no lineales. Finalmente, en 1989 se demuestra que MLP es un aproximador universal, es decir, es capaz de aproximar cualquier función continua. Hoy en día, es una de las arquitecturas más empleadas [66].

Un perceptrón multicapa es un tipo de red neuronal artificial constituida por una capa de entrada, una de salida y una o más capas ocultas. Sin embargo, se ha demostrado que, por lo general, basta con una sola capa oculta para obtener resultados óptimos, motivo por el que se opta en este trabajo por construir una red con solo una de ellas [67].

A continuación, se explica más detalladamente esta estructura [66]:

- **Capa de entrada**. Su función se basa en la recepción de las entradas al modelo y propagación de las mismas a la siguiente capa.
- **Capa de salida.** Devuelve el resultado de la red generada por cada patrón de entrada.
- **Capas ocultas**. Efectúan el procesamiento no lineal de las entradas de la red.

Para complementar esta descripción, se muestra la arquitectura clásica de un perceptrón multicapa constituido por una sola capa oculta (Figura 19):

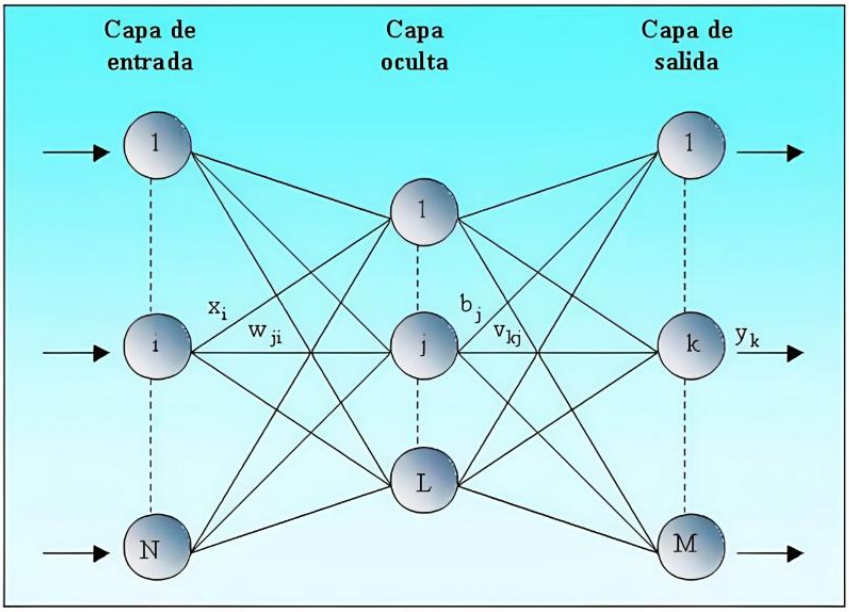

Figura 19. Arquitectura de perceptrón multicapa de una sola capa oculta [67].

x_i: entrada recibida por la neurona "i" de la capa de entrada.
w_{ji}: peso de conexión entre la neurona "i" de la capa de entrada y la neurona "j" de la capa oculta.
b_j: salida de la neurona oculta "j".
v_{kj}: peso de conexión entre la neurona "j" de la capa oculta y la neurona "k" de la capa de salida.
y_k: salida de la neurona "k" de la capa de salida.

Como puede observarse, la capa de entrada consta de *N* neuronas. Cada una de ellas recibe una entrada *x*, que es transferida a todas las neuronas que constituyen la capa oculta. En dicha capa, las entradas se combinan con los pesos *w* y, mediante la aplicación de una función de activación, se generan las salidas *b* de cada neurona de la capa oculta. Finalmente, estas salidas son transmitidas a la capa de salida, combinándose con los pesos *v*. De este modo, se obtiene la salida final del modelo.

Otra característica de estos modelos es que son redes *feedforward,* es decir, las conexiones entre las neuronas se establecen siempre desde una capa hacia las neuronas de la siguiente [67]. Además, no pueden existir conexiones entre neuronas que formen parte de la misma capa (conexiones laterales). Por tanto, el flujo de información en estos casos es unidireccional.

Respecto al cálculo que se efectúa en este tipo de redes, para cada neurona oculta se obtiene su entrada neta o total. Por ejemplo, para la neurona *j* de la capa oculta, esta entrada neta (net_j) se define como [67]:

$$net_j = \sum_{i=1}^{N} w_{ji} x_i + \theta_j \tag{4.2}$$

θ_j representa el sesgo, que se suma para ajustar la activación de la neurona. A raíz de esta entrada neta, es posible determinar la salida de esta misma neurona (b_j) implementando la función de activación *f* [67]:

$$b_j = f(net_j) \tag{4.3}$$

Análogamente, la entrada neta para la neurona k de la capa de salida (net_k) se rige por la siguiente ecuación [67]:

$$net_k = \sum_{j=1}^{L} v_{kj} b_j + \theta_k \tag{4.4}$$

Igual que en el caso anterior, θ_k constituye el sesgo. Por último, la salida de esta neurona (y_k) se calcula aplicando la función de transferencia a su entrada neta (net_k) [67]:

$$y_k = f(net_j) \tag{4.5}$$

Por tanto, cada neurona suma las entradas multiplicadas por los pesos y las ajusta mediante el sesgo. El resultado de esto es transformado por una función de activación, generando así una salida que se emite hacia la siguiente capa.

En este TFG, el modelo predictivo basado en el MLP fue implementado con las dos funciones de activación más empleadas en la actualidad: la tangente hiperbólica para la capa oculta y la función sigmoidal para la capa de salida. A continuación, se incluyen sus representaciones gráficas, expresiones matemáticas y una breve descripción de cada una de ellas:

- **Tangente hiperbólica**. Es muy parecida a la sigmoide, pero a diferencia de esta, devuelve valores entre -1 y 1 (Figura 20). Además, dicha salida está centrada en 0, lo que indica que las entradas negativas producen salidas negativas, mientras que las positivas generan salidas positivas. Una ventaja clave de esta función de activación es que puede proporcionar un entrenamiento más rápido y estable [68]. Todas estas prestaciones justifican su uso frecuente en las capas ocultas de redes neuronales y explican por qué se optó finalmente por aplicarla como función de transferencia en la capa oculta del modelo predictivo.

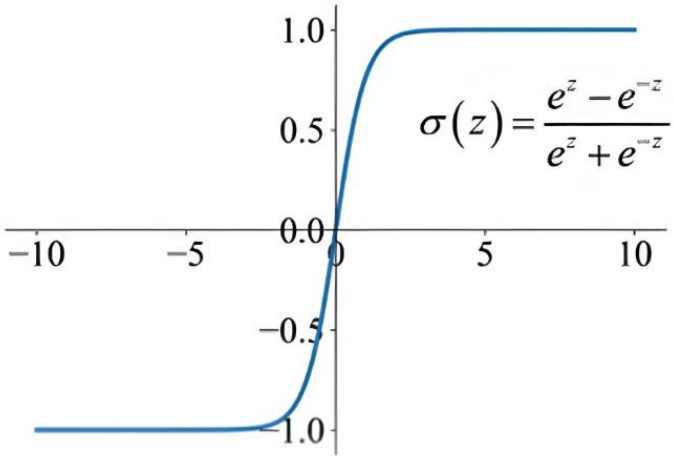

Figura 20. Gráfica de la función de activación de la tangente hiperbólica y su expresión matemática [69].

- **Función sigmoidal**. Devuelve valores entre 0 y 1, lo que la hace especialmente útil en problemas de clasificación binaria, como es el caso del presente estudio (reingreso vs. no reingreso) [69]. Este rasgo característico permite interpretar su salida como valores probabilísticos, motivo por el que se aplicó en la capa de salida del modelo diseñado. En la Figura 21, se expone el aspecto y expresión matemática de esta función:

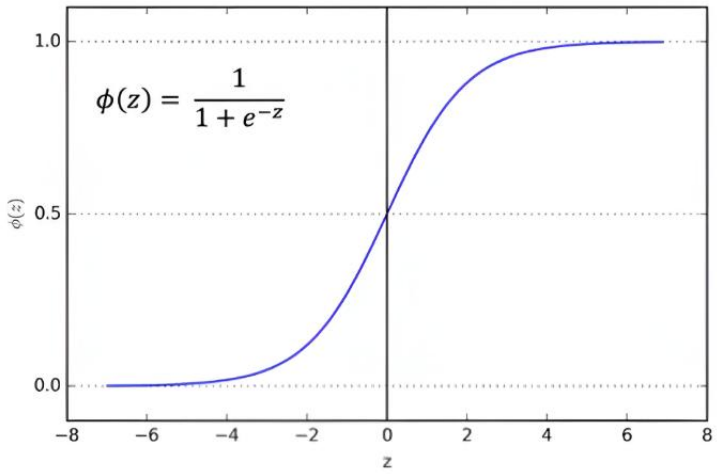

Figura 21. Gráfica de la función de activación sigmoidal y su expresión matemática [69].

Análogamente al modelo predictivo basado en *Random Forest*, se optimizaron los hiperparámetros de la red neuronal perceptrón multicapa. El enfoque se centró en ajustar el número de neuronas en la única capa oculta y el parámetro de regularización.

La estrategia consistió en entrenar la red probando diferentes combinaciones de estos hiperparámetros mediante un bucle anidado, empleando el conjunto *train2* para el entrenamiento (véase Figura 17). El rendimiento de cada una de estas combinaciones se calculó mediante la métrica *F1 score* en el conjunto de validación, seleccionando finalmente aquella que maximizase esta métrica.

Tras la obtención del número de neuronas y el parámetro de regularización óptimos, se procedió a ajustar el umbral de decisión, que por defecto es 0.5. La dinámica aplicada fue la misma que para optimizar los hiperparámetros, es decir, se consideró como mejor umbral aquel cuyo *F1 score* fuese superior al resto en el conjunto de validación.

Definidos los tres parámetros (número de neuronas en la capa oculta, regularización y umbral), el modelo predictivo fue reentrenado con el conjunto de entrenamiento *train1* y se evaluó en el conjunto *test*.

4.6. Parámetros y métricas de rendimiento para evaluación de modelos predictivos

La evaluación de los modelos predictivos binarios expuestos anteriormente (*Random Forest* y la red neuronal perceptrón multicapa) ha sido efectuada mediante una serie de métricas que cuantifican la capacidad de los mismos para discriminar adecuadamente entre pacientes que reingresarán y pacientes que no lo harán en los 30 días posteriores al alta. Todas las métricas aplicadas derivan de cálculos realizados a partir de los parámetros de la matriz de confusión.

La matriz de confusión, creada en 1904 por Karl Pearson, es una matriz cuadrada de *tamaño N x N*, donde N alude al número de clases de salida [70]. En este trabajo en particular, solo existen dos posibles clases: positiva (reingreso) o negativa (no reingreso). Por tanto, la dimensión de la matriz construida es de *2 x 2*.

En cuanto a su estructura, aunque el orden puede ser intercambiable, se ha confeccionado una matriz de confusión en la que las filas denotan la clase verdadera, mientras que las columnas representan la clase predicha [71]. A continuación, en la Tabla 10 se muestra el aspecto general de la misma:

Tabla 10. Estructura de una matriz de confusión de clasificación binaria.

CLASE DE DATOS		**Clase predicha (*predicted class*)**	
		0 = No reingreso	**1 = Reingreso**
Clase verdadera (*true class*)	**0 = No reingreso**	TN	FP
	1 = Reingreso	FN	TP

TN: verdaderos negativos; FP: falsos positivos; FN: falsos negativos; TP: verdaderos positivos

Sus componentes son:

- **Verdaderos negativos (TN):** número de pacientes predichos por el modelo como no reingreso y que realmente no reingresan en los 30 días posteriores al alta.

- **Falsos positivos (FP):** número de pacientes que el modelo predice que serán readmitidos por exacerbación de EPOC, pero que finalmente acaban no reingresando en los 30 días posteriores al alta.

- **Falsos negativos (FN):** número de pacientes predichos por el modelo como no reingreso, pero que sí reingresan en los 30 días posteriores al alta.

- **Verdaderos positivos (TP):** número de pacientes que el modelo predice que reingresarán y que, efectivamente son readmitidos por exacerbación de EPOC en los 30 días posteriores al alta.

Las métricas de rendimiento, como se ha indicado previamente, derivan de los elementos de la matriz de confusión definidos. Estas son: sensibilidad, especificidad, valor predictivo positivo, valor predictivo negativo, razón de verosimilitud positiva, razón de verosimilitud negativa, precisión, *F1 score* y área bajo la curva ROC (AUC*)*.

- **Sensibilidad *(Se)***. También conocida como *recall*, es la fracción de casos positivos (reingresos) predichos como positivos [72]:

$$Se = \frac{TP}{TP+FN} \tag{4.6}$$

- **Especificidad (*Sp*).** Fracción de casos negativos (no reingresos) predichos como negativos [72]:

$$Sp = \frac{TN}{TN+FP} \tag{4.7}$$

- **Valor predictivo positivo (*PPV*)**. Fracción de casos positivos reales (reingresos) respecto al total de casos que el modelo predijo como positivos [72]:

$$PPV = \frac{TP}{TP+FP} \tag{4.8}$$

- **Valor predictivo negativo (*NPV*)**. Fracción de casos negativos verdaderos (no reingresos) respecto al total de casos que el modelo predijo como negativos [72]:

$$NPV = \frac{TN}{TN+FN} \tag{4.9}$$

- **Razón de verosimilitud positiva (*positive likelihood ratio, LR+*).** Describe cuánto más probable es que el test bajo estudio sea positivo al aplicarse sobre un paciente realmente positivo o patológico que al aplicarse sobre uno negativo [73]:

$$LR^{+} = \frac{Se}{1-Sp} \tag{4.10}$$

- **Razón de verosimilitud negativa (*negative likelihood ratio*, *LR-*).** Describe cuánto más probable es que el test bajo estudio sea negativo al aplicarse sobre un sujeto realmente positivo o patológico que al aplicarse sobre uno negativo/no-patológico [73]:

$$LR^{-} = \frac{1-Se}{Sp} \tag{4.11}$$

- **Precisión *(Acc)*.** También conocida como *accuracy*, es la proporción de predicciones correctas respecto al total de predicciones efectuadas [70]:

$$Acc = \frac{TP+TN}{TP+FN+TN+FP} \tag{4.12}$$

- ***F1 score*.** Media armónica del valor predictivo positivo y la sensibilidad. Como se ha adelantado en anteriores apartados, esta fue la métrica principal para la optimización de los hiperparámetros de los modelos predictivos propuestos. El motivo de esta decisión se basa en la capacidad de *F1 score* para equilibrar sensibilidad y el valor predictivo positivo (PPV). Asimismo, a pesar de que la precisión suele ser la métrica habitual aplicada en estos contextos, *F1 score* resulta más idónea cuando se dispone de bases de datos con clases desbalanceadas, como es el caso de este trabajo [70]. Se define como [72]:

$$F1\ score = 2 \cdot \frac{PPV \cdot Se}{PPV+Se} \tag{4.13}$$

- **Curva ROC y área bajo la curva ROC (AUC):**

  La curva ROC (*Receiver Operating Characteristic*) es una gráfica que representa la tasa de verdaderos positivos (TPR) o sensibilidad*,* frente a la tasa de falsos positivos (FPR), calculada como *1 – especificidad*. Permite observar el rendimiento del modelo en todos los umbrales de clasificación trazando los valores de FPR y TPR para cada umbral. Es por ello que resulta una herramienta muy útil para establecer un umbral de clasificación. El aspecto ideal de una curva ROC se caracteriza por poseer una tasa de verdaderos positivos del 100% y una tasa de falsos positivos del 0% [70].

  Un parámetro importante que puede extraerse de la misma es el área bajo la curva ROC (AUC), comúnmente aplicada para evaluar modelos predictivos de clasificación binaria. Esta métrica mide el área total bajo la curva y es capaz de predecir la calidad independientemente del umbral de decisión establecido. Su valor oscila entre 0.5 y 1, siendo 0.5 un comportamiento aleatorio y 1 un indicativo de que es completamente correcto [70].

  En la Figura 22, se muestra un ejemplo de una curva ROC, ilustrando diferentes rendimientos hipotéticos mediante distintos colores [70].

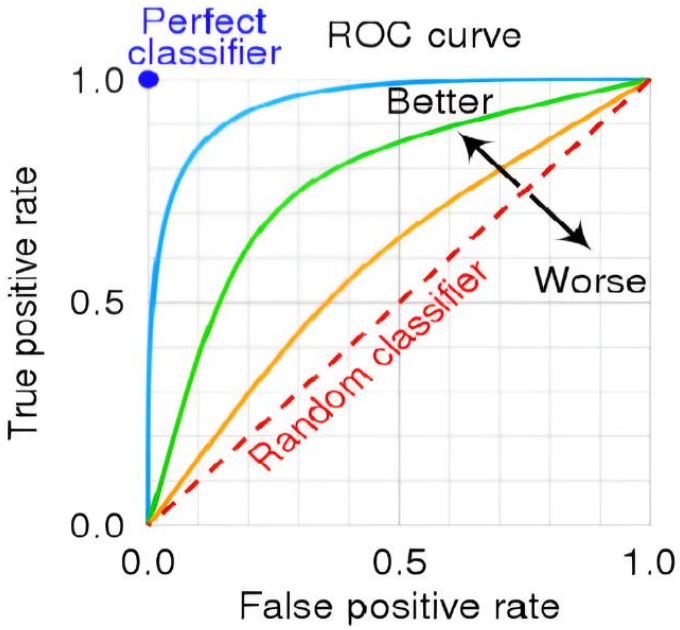

Figura 22. Aspecto de la curva ROC y diferentes escenarios posibles [70].

Los escenarios mostrados se detallan a continuación:

- Gráfica roja discontinua: dada su coincidencia con la línea diagonal, el rendimiento que refleja es característico de un clasificador aleatorio, es decir, el modelo predictivo correspondiente no tiene capacidad para diferenciar entre pacientes que reingresarán y aquellos que no lo harán. Además, su AUC es de aproximadamente 0.5 y valores más bajos supondrían modelos predictivos de muy baja calidad.

- Gráfica naranja continua: se encuentra ligeramente por encima de la diagonal, por lo que es indicativo de un rendimiento superior al de un clasificador aleatorio, pero aún algo limitado.

- Gráfica verde continua: su rápido crecimiento y mayor proximidad a la esquina superior izquierda ilustran un modelo predictivo con una capacidad predictiva destacable. Asimismo, el valor de AUC en este caso se aproxima a 1 (valor ideal).

- Gráfica azul continua: representa el rendimiento predictivo de un modelo ideal y una *AUC = 1.* No obstante, este escenario es difícil de alcanzar, puesto que no puede ignorarse la naturaleza variable de los datos clínicos.

4.7. Análisis estadístico

4.7.1. Variables continuas: prueba *U* de Mann-Whitney

Para poder establecer una comparativa de las variables continuas entre los grupos reingreso (clase 1) y no reingreso (clase 0), se ha aplicado la **prueba *U* de Mann-Whitney** considerando un nivel de significancia estadística de α = 0.05. Esta prueba permite determinar la existencia de diferencias estadísticamente significativas entre dos grupos independientes [74]. Además, dado que es una técnica no paramétrica, una de sus principales características es que es de "distribución libre", es decir, no asume una distribución concreta sobre la población. Asimismo, su aplicación resulta idónea cuando se trata de muestras reducidas, hecho que también lo hace útil en el contexto médico y, por tanto, en este trabajo [75].

Las hipótesis a contrastar en esta prueba son [74]:

- **Hipótesis nula (H_0):** no existe diferencia entre los dos grupos independientes.
- **Hipótesis alternativa (H_1):** existe una diferencia entre los dos grupos independientes.

Para efectuar su cálculo, en primer lugar, deben combinarse todas las observaciones de la variable concreta a analizar de ambos conjuntos de datos. De esta forma, se dispone de un único grupo de tamaño *N* (suma total del número de muestras) sobre el que se asigna un rango en sentido creciente desde 1 hasta *N* [75], [76].

Una vez establecido el *ranking,* se efectúa la suma de los rangos para cada grupo, calculándose así los parámetros T_a y T_b. Siendo n_a y n_b el tamaño muestral de cada uno de los conjuntos, el estadístico *U* se obtiene de la siguiente manera [76]:

- Si $n_a > n_b$:

$$U = T_a - \frac{n_a(n_a+1)}{2} \tag{4.14}$$

- Si $n_a < n_b$:

$$U = T_b - \frac{n_b(n_b+1)}{2} \tag{4.15}$$

Finalmente, este estadístico *U* es comparado con un valor crítico para poder dictaminar si se rechaza H_0. Si *U* es mayor que este, se rechaza la hipótesis. Normalmente, el nivel de significancia aplicado es del 0.05.

4.7.2. Variables categóricas: test exacto de Fisher

El test exacto de Fisher es una herramienta que permite determinar si existe una diferencia estadísticamente significativa entre dos variables categóricas [77]. Además, su precisión aumenta ante un tamaño bajo de observaciones [78], hecho que lo hace idóneo para el ámbito médico y este TFG. Análogamente a la prueba *U* de Mann-Whitney, se ha aplicado un nivel de significancia estadística de $\alpha = 0.05$.

Las hipótesis contrastadas en este test son [77]:

- **Hipótesis nula (H_0):** no existe asociación entre los grupos reingreso y no reingreso, es decir, no hay diferencias significativas.
- **Hipótesis alternativa (H_1):** existe asociación entre los grupos.

Para realizar el test, los datos recopilados se disponen en una tabla de contingencia, con al menos dos filas (una categoría de esa variable y las demás combinadas) y dos columnas (reingreso y no reingreso). Esta tabla contiene las frecuencias de aparición de cada combinación de categoría (fila) y grupo (columna) [79]. En este trabajo, se genera una tabla de contingencia para cada categoría de una variable y se compara con las otras categorías de dicha variable juntas. Para ilustrar bien este concepto y ofrecer una mayor claridad en la explicación, se presenta en la Tabla 11 un ejemplo de la estructura de una tabla de contingencia con una fila y columna de totales:

Tabla 11. Ejemplo de estructura de una tabla de contingencia para una variable categórica.

	Reingreso	**No reingreso**	**Total**
Categoría específica	*a*	*b*	*a + b*
Resto de categorías	*c*	*d*	*c + d*
Total	*a + c*	*b + d*	*n = a + b + c + d*

- *a*: pacientes que reingresaron y pertenecen a la categoría específica.
- *b*: pacientes que no reingresaron y pertenecen a la categoría específica.
- *c*: pacientes que reingresaron y pertenecen al resto de categorías.
- *d*: pacientes que no reingresaron y pertenecen al resto de categorías.
- *n*: total de observaciones.

Con ello, es posible calcular la probabilidad exacta de obtener la distribución observada en la tabla atendiendo a la siguiente fórmula [78]:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \tag{4.16}$$

Tras esto, se calcula el p-valor sumando la probabilidad de la tabla de contingencia y de todas aquellas cuyos sumatorios en filas y columnas sean los mismos, aunque con diferencias más notorias entre reingreso y no reingreso.

CAPÍTULO 5. RESULTADOS

El presente capítulo expone los hallazgos derivados del proceso metodológico descrito en la anterior sección. Entre estos resultados se incluye el análisis de la población bajo estudio para obtener los pacientes que finalmente formaron parte del trabajo, la descripción estadística de las variables disponibles, el análisis de datos perdidos por paciente y por variable, la selección final de características predictoras que constituirán las entradas a los modelos predictivos, la optimización de los hiperparámetros en los modelos de *Random Forest* y red neuronal perceptrón multicapa, y la evaluación de su capacidad predictiva mediante las métricas pertinentes. Todos estos resultados sientan las bases para la posterior discusión que se abordará en el siguiente capítulo.

5.1. Población bajo estudio

En este TFG se dispone de dos poblaciones: (i) retrospectiva, para diseño y validación interna; y (ii) prospectiva, para validación temporal independiente. A continuación, se analiza el flujo de pacientes y sus características de ambas por separado, ya que son independientes entre sí.

- **Población retrospectiva.** Un total de 246 pacientes fueron fieles a los criterios de inclusión establecidos y firmaron asimismo el consentimiento informado. Sin embargo, de estos, 3 sujetos fueron excluidos, ya que representaban pacientes cuya alta fue motivada por su fallecimiento. Por consiguiente, no pudo efectuarse el seguimiento estipulado durante el periodo de 30 días posteriores al alta. Este descarte resulta finalmente en un tamaño muestral de 243 registros, siendo el 76.54% hombres (186 en cifras absolutas) y el 23.46% mujeres (57 en cifras absolutas). Además, 42 son reingresos (clase positiva) y 201 no reingresos (clase negativa), con una edad media de 73.48 años. El número medio de días transcurridos desde el alta al reingreso fue de 15.43 días, falleciendo el 4.76% de sujetos en el transcurso de su rehospitalización.

  A continuación, en la Figura 23, se expone el diagrama de flujo descriptivo de la población retrospectiva final:

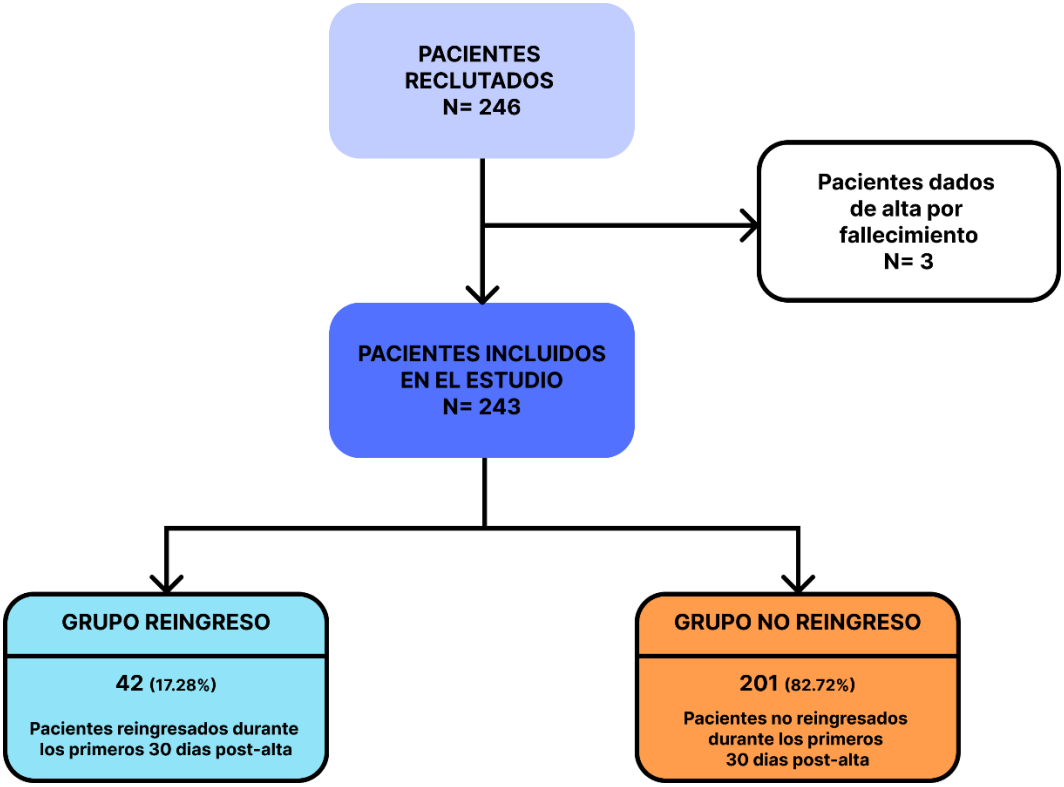

Figura 23. Diagrama de flujo de pacientes que forman parte de la población retrospectiva.

- **Población prospectiva.** Un total de 75 pacientes cumplieron los criterios de inclusión acordados (los mismos que en la base de datos retrospectiva) y firmaron el consentimiento informado. No obstante, de ellos, 1 paciente fue registrado como alta por fallecimiento. De los sujetos restantes, en 64 hubo dificultades en la adquisición de algunas variables. Finalmente, pudieron ser analizados un total de 10 pacientes (8 hombres y 2 mujeres). Además, 1 reingresó (clase positiva) y 9 no reingresaron (clase negativa).

  A continuación, en la Figura 24, se presenta el diagrama de flujo descriptivo de la población prospectiva final:

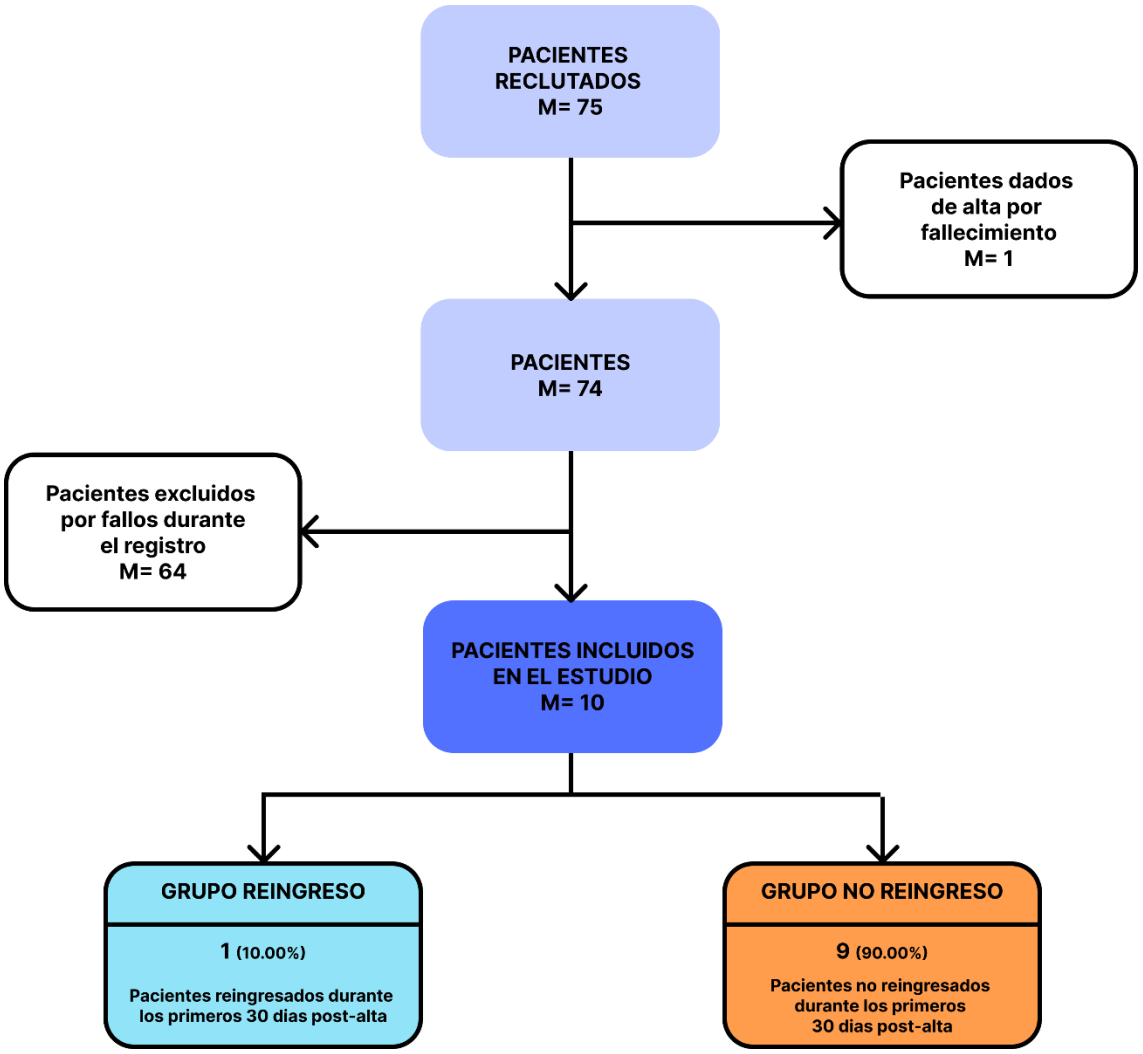

Figura 24. Diagrama de flujo de pacientes que forman parte de la población prospectiva.

5.2. Análisis de datos perdidos

A continuación, se exponen los resultados del tratamiento de datos perdidos (*NaNs*). Esta etapa resulta fundamental, ya que, como se expresó en el capítulo *Metodología,* un adecuado abordaje evita comprometer la validez y robustez de los modelos predictivos diseñados.

En primer lugar, como se indicó previamente, se efectuó una exploración inicial de los datos faltantes por variable y paciente antes de la aplicación de cualquier estrategia o filtrado. Esto permitió conocer aquellos pacientes y variables cuya proporción de datos perdidos era elevada.

En la Tabla 12, se recoge el porcentaje de datos perdidos por variable, apreciándose diversas características cuyo valor superó el umbral establecido (10%). De hecho, la elección de este valor como umbral para la eliminación de variables tiene su origen en esta exploración inicial. A pesar de que se identificó una variable con un porcentaje de *NaNs* próximo al 10% (en concreto 9.88%),

la variable consecutiva en orden creciente de datos perdidos presentaba un 28.81% de *NaNs*. Por consiguiente, se optó por determinar un valor del 10% como límite razonable con el que se pretendía conservar el máximo número de variables posible, a la vez que no interferir en la calidad de los datos.

Para favorecer una mejor comprensión y claridad de los resultados, se destacan en rojo las variables con un porcentaje superior al umbral. Por otra parte, en amarillo se denotan aquellas con nula o mínima variabilidad o representatividad, es decir, cuando la totalidad de los pacientes, o todos menos uno de ellos, pertenezcan a la misma clase.

Tabla 12 - I. Exploración de datos perdidos por variable antes del filtrado.

Variable	% *NaNs*	Variable	% *NaNs*	Variable	% *NaNs*
Edad	0.00	**Medicación de rescate basal LABA**	0.00	**Complicación: sepsis**	0.00
Sexo	0.00	**Medicación de rescate basal LAMA**	0.00	**Complicación: insuficiencia respiratoria**	0.00
Peso	0.00	**Medicación de rescate basal corticoides**	0.00	**Medicación inhaladora al ingreso, ninguna**	0.00
Altura	0.00	**Corticoterapia sistémica basal continua**	0.00	**Medicación inhaladora al ingreso, SABA**	0.00
Índice de Masa Corporal (IMC)	0.00	**Teofilinas en estado basal**	0.00	**Medicación inhaladora al ingreso, SAMA**	0.00
Procedencia	0.00	**IDFE4 en estado basal**	0.00	**Medicación inhaladora al ingreso, LABA**	010
Estado civil	0.00	**Mucolíticos en estado basal**	0.00	**Medicación inhaladora al ingreso, LAMA**	0.00
Hogar	0.00	**Antibióticos en estado basal**	0.00	**Medicación inhaladora al ingreso, corticoides**	0.00
Actividad laboral	0.00	**Test mMRC**	0.00	**Corticoterapia sistémica al ingreso**	0.00
Movilidad	0.41	**Test CAT**	2.06	**Ingreso en UVI**	0.00
Cuidador	0.41	**Test TAI_Ade**	1.65	**Oxigenoterapia al ingreso**	0.00
Test Barthel	0.00	**Test_TAI_I_Err**	1.65	**Teofilinas al ingreso**	0.00
Tabaquismo	0.00	**Test_TAI_I_Del**	1.65	**IDFE4 al ingreso**	0.00
Índice tabáquico	0.00	**Test_TAI_I_Inc**	1.65	**Mucolíticos al ingreso**	0.00
Alcohol	0.41	**Test EuroQol-5D: movilidad**	0.00	**Ventilación no invasiva al ingreso**	0.00

Test_TAI_Ade: test de TAI, nivel de adhesión; Test_TAI_I_Err: test de TAI, incumplimiento errático; Test_TAI_I_Del: test de TAI, incumplimiento deliberado; Test_TAI_I_Inc: test de TAI, incumplimiento inconsciente.

Tabla 12 - II (cont.) Exploración de datos perdidos por variable antes del filtrado.

Variable	% *NaNs*	Variable	% *NaNs*	Variable	% *NaNs*
Gramos diarios de alcohol	76.54	Test_E5D_CPe	0.00	Ventilación invasiva al ingreso	0.00
Grupo de riesgo clínico	0.00	Test_E5D_ACo	0.00	Días de ventilación invasiva	99.59
Tratamiento anticoagulación	0.00	Test_E5D_Dol	0.00	Antibióticos al ingreso	0.00
Tratamiento antiagregación	0.00	Test_E5D_Dep	0.00	Tensión sistólica al alta	28.81
Vacuna antigripal	0.00	Número de días ingresado	0.00	Tensión diastólica al alta	28.81
Vacuna antineumocócica	0.00	Tensión sistólica al ingreso	2.06	Saturación basal al alta	65.02
Hipertensión arterial	0.00	Tensión diastólica al ingreso	2.06	Saturación con O_2 suplementario al alta	57.20
Diabetes mellitus	0.00	Saturación basal al ingreso	32.10	Litros de O_2 suplementario al alta	58.85
Dislipemia	0.00	Saturación con O_2 suplementario al ingreso	56.79	Frecuencia cardíaca al alta	30.86
Cardiopatía isquémica	0.00	Litros de O_2 suplementario al ingreso	67.49	Frecuencia respiratoria al alta	94.24
Insuficiencia cardíaca	0.00	Frecuencia cardíaca al ingreso	2.47	pH al alta	60.49
Insuficiencia respiratoria	0.00	Frecuencia respiratoria al ingreso	41.56	PCO_2 al alta	60.49
Bronquiectasias	0.00	pH al ingreso	7.00	PO_2 al alta	60.49
Revascularización	0.00	PCO_2 al ingreso	7.00	HCO3 al alta	60.49
Taquiarritmia	0.00	PO_2 al ingreso	7.00	A_Gas_A_a	73.25
Accidente cerebrovascular	0.00	HCO_3 al ingreso	7.41	FiO2 al alta	69.14
Demencia	0.00	Gradiente arterio-alveolar al ingreso	53.50	FVC al alta	97.12
Neoplasia pulmonar	0.00	FiO2 al ingreso	39.51	FVC al alta (% teórico)	97.12
Otras neoplasias	0.00	Leucocitos al ingreso	1.65	FVC al alta (z score)	98.35
Enfermedad renal	0.00	Neutrófilos al ingreso	1.65	FEV1 al alta	97.12
Osteoporosis	0.00	Neutrófilos al ingreso (%)	1.65	FEV1_al alta (% teórico)	97.12
Ansiedad	0.00	Eosinófilos al ingreso	1.65	FEV1_al ata (z score)	98.35
Depresión	0.00	Eosinófilos al ingreso (%)	1.65	FEV1/FVC al alta	97.12
Anemia	0.00	Aumento tos	0.00	FEV1/FVC al alta (% teórico)	98.35
Tromboembolismo pulmonar	0.00	Aumento disnea	0.00	FEV1/FVC al alta (z score)	98.35

Test_E5D_CPe: test EuroQoL-5D, cuidado personal; Test_E5D_Aco: test EuroQoL-5D, actividades cotidianas; Test_E5D_Dol: test EuroQoL-5D, dolor; Test_E5D_Dep: test EuroQoL-5D, depresión.

Tabla 12 - III (cont.) Exploración de datos perdidos por variable antes del filtrado.

Variable	% NaNs	Variable	% NaNs	Variable	% NaNs
Síndrome apnea-hipopnea	0.00	Aumento expectoración	0.00	Leucocitos al alta	55.14
Test Charlson	0.00	Purulencia esputo	0.00	Neutrófilos al alta	55.14
FVC basal	1.65	Dolor torácico	0.00	Neutrófilos al alta (%)	55.14
FVC basal (% teórico)	1.65	Fiebre	0.00	Eosinófilos al alta	55.14
FVC basal (z score)	1.65	Número de días de síntomas	9.88	Eosinófilos al alta (%)	55.14
FEV1 basal	1.23	Uso musculatura accesoria	0.00	Oxigenoterapia domiciliaria al alta	0.00
FEV1 basal (% teórico)	1.23	Movimientos torácicos	0.00	Ventilación no invasiva al alta	0.00
FEV1 basal (z score)	1.23	Cianosis	0.00	Medicación inhaladora al alta, ninguna	0.00
FEV1/FVC basal	1.23	Edemas periféricos	0.00	Medicación inhaladora al alta, SABA	0.00
FEV1/FVC basal (% teórico)	1.23	Inestabilidad hemodinámica	0.00	medicación inhaladora al alta, SAMA	0.00
FEV1/FVC basal (z score)	1.23	Deterioro mental	0.00	Medicación inhaladora al alta, LABA	0.00
Estratificación del riesgo	0.00	Parada respiratoria	0.00	Medicación inhaladora al alta, LAMA	0.00
Fenotipo	0.00	Disnea	0.00	Medicación inhaladora al alta, corticoides;	0.00
Grado de obstrucción	1.23	Causa infecciosa	0.00	Medicación de rescate al alta, ninguna	0.00
Gold	0.00	Causa bacteriana	0.00	Medicación de rescate al alta, SABA	0.00
Ingresos por agudización (año previo)	0.00	Causa vírica	0.00	Medicación de rescate al alta, SAMA	0.00
Oxigenoterapia domiciliaria basal	0.00	Microorganismos resistentes	0.00	Medicación de rescate al alta, LAMA	0.00
Ventilación no invasiva basal	0.00	Cultivo germen	49.38	Medicación de rescate al alta, corticoides	0.00
Medicación inhaladora basal, ninguna	0.00	Complicación: arritmias	0.00	Corticoterapia sistémica al alta	0.00
Medicación inhaladora basal, SABA	0.00	Complicación: insuficiencia cardíaca	0.00	Antibióticos al alta	0.00

Tabla 12 - IV (cont.) Exploración de datos perdidos por variable antes del filtrado.

Variable	% NaNs	Variable	% NaNs	Variable	% NaNs
Medicación inhaladora basal, SAMA	0.00	**Complicación: cardiopatía isquémica**	0.00	**Teofilinas al alta**	0.00
Medicación inhaladora basal, LABA	0.00	**Complicación: derrame pleural**	0.00	**IDFE4 al alta**	0.00
Medicación inhaladora basal, LAMA	0.00	**Complicación: neumonía**	0.00	**Mucolíticos al alta**	0.00
Medicación inhaladora basal, corticoides	0.00	**Complicación: tromboembolismo pulmonar**	0.00	**Rehabilitación respiratoria al alta**	0.00
Medicación de rescate basal, ninguna	0.00	**Complicación: neumotórax**	0.00		
Medicación de rescate basal, SABA	0.00	**Complicación: síndrome de distrés respiratorio agudo**	0.00		
Medicación de rescate basal, SAMA	0.00	**Medicación de rescate al alta, LABA**	0.00		

En esta Tabla 12, las variables marcadas suponen un total de **42** características, que serán eliminadas por superar el umbral del 10% de datos ausentes (**36** variables, en rojo) o por implicar una nula o mínima variabilidad/representatividad de una de las clases (**6** variables, en amarillo). Las restantes constituyen el conjunto de características sobre las que se aplicará el proceso de selección de variables posterior una vez imputadas. En el caso de los pacientes, se calculó el porcentaje de datos faltantes, puesto que la supresión de variables influye directamente en la cantidad de *NaNs* por sujeto. Esto último es recogido en la Tabla 13.

Dado que la pérdida máxima de datos por paciente fue del 7.59%, no se suprimió ningún sujeto, en contraste con lo efectuado con las variables. De esta manera, se priorizó mantener el mayor número de pacientes en el estudio, ya que el tamaño de la población recogida es reducido.

Además, en la base de datos prospectiva se detectó una variable con un significativo porcentaje de datos perdidos. Esto ocasionó el descarte de un elevado número de pacientes, siendo excluidos 36 sujetos (48% de la cohorte inicial compuesta por 75 individuos).

Tabla 13 - I. Exploración de datos perdidos por paciente después del filtrado.

Paciente	% *NaNs*	Paciente	% *NaNs*	Paciente	% *NaNs*
1	2.53	**82**	0.00	**163**	0.00
2	0.00	**83**	0.00	**164**	0.00
3	0.00	**84**	0.00	**165**	0.00
4	0.00	**85**	0.00	**166**	0.00
5	0.00	**86**	0.00	**167**	0.00
6	2.53	**87**	0.00	**168**	2.53
7	0.00	**88**	0.00	**169**	0.00
8	0.00	**89**	0.00	**170**	4.43
9	0.00	**90**	0.00	**171**	0.63
10	0.00	**91**	0.00	**172**	0.63
11	0.00	**92**	0.00	**173**	1.90
12	0.00	**93**	0.00	**174**	0.63
13	0.00	**94**	0.00	**175**	0.00
14	0.00	**95**	0.00	**176**	0.63
15	0.00	**96**	0.00	**177**	0.00
16	2.53	**97**	0.00	**178**	0.00
17	0.00	**98**	0.00	**179**	0.63
18	0.00	**99**	0.00	**180**	0.00
19	0.00	**100**	0.00	**181**	0.00
20	0.00	**101**	0.00	**182**	0.00
21	0.00	**102**	0.00	**183**	0.63
22	0.00	**103**	2.53	**184**	3.80
23	0.00	**104**	7.59	**185**	0.00
24	0.00	**105**	0.00	**186**	0.63
25	0.00	**106**	0.00	**187**	0.00
26	2.53	**107**	0.00	**188**	0.63
27	0.00	**108**	0.00	**189**	0.00
28	0.00	**109**	0.00	**190**	0.00
29	0.00	**110**	0.00	**191**	0.00
30	0.00	**111**	0.00	**192**	0.00
31	0.00	**112**	0.00	**193**	0.00
32	0.63	**113**	0.00	**194**	0.00
33	0.00	**114**	1.90	**195**	0.00
34	0.00	**115**	2.53	**196**	0.63
35	2.53	**116**	0.00	**197**	0.63
36	0.00	**117**	1.90	**198**	0.00
37	0.00	**118**	6.33	**199**	0.00
38	0.00	**119**	0.00	**200**	0.00
39	0.00	**120**	2.53	**201**	0.63
40	0.00	**121**	0.00	**202**	0.00
41	0.00	**122**	0.63	**203**	3.16
42	0.00	**123**	0.63	**204**	0.00
43	2.53	**124**	3.80	**205**	0.63
44	0.00	**125**	0.00	**206**	0.00
45	2.53	**126**	0.00	**207**	0.00

Tabla 13 - II (cont.) Exploración de datos perdidos por paciente después del filtrado.

Paciente	% *NaNs*	Paciente	% *NaNs*	Paciente	% *NaNs*
46	0.00	**127**	0.63	**208**	0.00
47	0.00	**128**	1.90	**209**	0.00
48	2.53	**129**	0.00	**210**	0.00
49	0.00	**130**	0.63	**211**	0.00
50	0.00	**131**	0.00	**212**	0.00
51	0.00	**132**	0.00	**213**	0.00
52	0.00	**133**	0.00	**214**	0.63
53	0.00	**134**	0.00	**215**	0.00
54	0.00	**135**	0.00	**216**	3.16
55	0.00	**136**	0.00	**217**	0.00
56	0.00	**137**	0.00	**218**	0.00
57	0.00	**138**	0.00	**219**	0.63
58	0.00	**139**	0.00	**220**	0.63
59	0.00	**140**	0.00	**221**	0.00
60	6.33	**141**	0.00	**222**	0.00
61	0.00	**142**	0.00	**223**	0.00
62	0.00	**143**	0.00	**224**	2.53
63	0.00	**144**	0.00	**225**	0.00
64	1.90	**145**	0.00	**226**	0.00
65	0.00	**146**	0.00	**227**	0.00
66	0.00	**147**	0.00	**228**	2.53
67	0.00	**148**	0.00	**229**	0.00
68	0.00	**149**	1.27	**230**	0.00
69	0.00	**150**	0.63	**231**	0.00
70	6.33	**151**	0.00	**232**	0.63
71	0.00	**152**	2.53	**233**	0.63
72	2.53	**153**	0.00	**234**	0.00
73	0.00	**154**	0.00	**235**	2.53
74	0.00	**155**	0.00	**236**	0.00
75	0.00	**156**	0.00	**237**	0.63
76	0.00	**157**	0.00	**238**	0.63
77	0.00	**158**	0.00	**239**	0.00
78	0.00	**159**	0.00	**240**	0.00
79	3.16	**160**	0.00	**241**	0.63
80	0.00	**161**	0.00	**242**	0.00
81	0.00	**162**	0.00	**243**	0.00

5.3. Análisis descriptivo de la base de datos

5.3.1. Base de datos retrospectiva

En la presente sección, se expone la caracterización estadística de las variables que conforman la base de datos principal (retrospectiva). Dada la naturaleza mixta propia de estas características, se ha optado por su agrupación en tablas (Tabla 14-27) según su temática,

manteniendo la clasificación que se indicó en el capítulo *Sujetos y Variables de Estudio*. A continuación de cada una de ellas, se exhiben los diagramas de cajas o *boxplots* correspondientes a aquellas variables continuas que presentan diferencias estadísticamente significativas entre las clases. La inclusión de los mismos es meramente informativa, ya que en el contexto de este trabajo no resulta razonable la identificación de *outliers.* Los valores obtenidos son un reflejo del grado de afectación de la enfermedad sobre el estado del paciente, por lo que la variabilidad mostrada en ellos es únicamente una evidencia de la heterogeneidad de la patología.

En relación con los **datos sociodemográficos y antropométricos** de los pacientes, la Tabla 14 muestra diferencias significativas entre clases en las variables *Hogar* y *Cuidador*. Respecto a la primera, la residencia resultó ser predominante en los reingresos. Por otra parte, la ausencia de cuidador fue más común en la clase negativa que en la positiva.

Tabla 14 - I. Caracterización de los datos sociodemográficos y antropométricos para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Edad	74.0 (66.0, 81.0)	74.0 (66.0, 81.0)	74.5 (67.0, 82.0)	0.4981
Sexo				
Mujer	57 (23.5%)	48 (23.9%)	9 (21.4%)	0.8427
Hombre	186 (76.5%)	153 (76.1%)	33 (78.6%)	
Peso	73.0 (63.0, 81.0)	73.0 (63.0, 83.0)	74.0 (64.0, 80.0)	0.9144
Altura	165.0 (160.0, 170.0)	165.0 (160.0, 170.0)	165.0 (158.0, 171.0)	0.7285
IMC	26.6 (24.0, 29.9)	26.5 (24.0, 30.1)	27.4 (24.4, 29.4)	0.6922
Procedencia				
Rural	78 (32.1%)	69 (34.3%)	9 (21.4%)	0.1451
Urbana	165 (67.9%)	132 (65.7%)	33 (78.6%)	
Estado civil				
Soltero	31 (12.8%)	29 (14.4%)	2 (4.8%)	0.1247
Casado	155 (63.8%)	129 (64.2%)	26 (61.9%)	0.8602
Viudo	39 (16.0%)	28 (13.9%)	11 (26.2%)	0.0634
Separado	18 (7.4%)	15 (7.5%)	3 (7.1%)	1.0000
Hogar				
Residencia	11 (4.5%)	6 (3.0%)	5 (11.9%)	**0.0253**
Solo	43 (17.7%)	38 (18.9%)	5 (11.9%)	0.3751
Familiares	189 (77.8%)	157 (78.1%)	32 (76.2%)	0.8387
Estudios				
Sin estudios	52 (21.4%)	44 (21.9%)	8 (19.0%)	0.8367
Primarios	153 (63.0%)	127 (63.2%)	26 (61.9%)	0.8626
Secundarios	25 (10.3%)	20 (10.0%)	5 (11.9%)	0.7795
Universitarios	13 (5.3%)	10 (5.0%)	3 (7.1%)	0.4754

Los datos se expresan en forma de mediana y rango intercuartil (IQR) para las variables continuas, y en número y porcentaje para las variables categóricas. IMC: Índice de Masa Corporal.

Tabla 14 – II (cont.) Caracterización de los datos sociodemográficos y antropométricos para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Actividad laboral				
Activo	8 (3.3%)	8 (4.0%)	0 (0.0%)	0.3572
Jubilado	215 (88.5%)	178 (88.6%)	37 (88.1%)	1.0000
Incapacitado	20 (8.2%)	15 (7.5%)	5 (11.9%)	0.3554
Baja laboral	0(0.0%)	0(0.0%)	0(0.0%)	~
Movilidad				
Cama-sillón	18 (7.4%)	15 (7.5%)	3 (7.1%)	1.0000
Paseo en domicilio	37 (15.2%)	27 (13.4%)	10 (23.8%)	0.1001
Salir a la calle	188 (77.4%)	159 (79.1%)	29 (69.0%)	0.1608
Cuidador				
No	112 (46.1%)	100 (49.8%)	12 (28.6%)	**0.0166**
Familiar	60 (24.7%)	50 (24.9%)	10 (23.8%)	1.0000
Cónyuge	53 (21.8%)	39 (19.4%)	14 (33.3%)	0.0631
Profesional	18 (7.4%)	12 (6.0%)	6 (14.3%)	0.0965

Los datos se expresan en forma de número y porcentaje para las variables categóricas.

Por otra parte, no se observaron diferencias significativas entre los grupos de estudio para las variables recogidas en la Tabla 15 relacionadas con los **hábitos** de los pacientes (*Tabaquismo, índice Tabáquico y Alcohol),* y las recogidas en la Tabla 16 relacionadas con los **datos clínicos** de los sujetos (*Grupo de riesgo clínico, Anticoagulación, Antiagregación, Vacuna antigripal en el año anterior y Vacuna antinemocócica*).

Tabla 15. Caracterización de los hábitos del paciente para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Tabaquismo				
No	6 (2.5%)	4 (2.0%)	2 (4.8%)	0.2775
Fumador	62 (25.5%)	53 (26.4%)	9 (21.4%)	0.5647
Ex fumador	175 (72.0%)	144 (71.6%)	31 (73.8%)	0.8518
Índice tabáquico	50.0 (30.0, 78.0)	52.0 (30.0, 78.5)	48.0 (33.0, 78.0)	0.8299
Alcohol				
No bebedor	184 (75.7%)	150 (74.6%)	34 (81.0%)	0.4352
Bebedor	59 (24.3%)	51 (25.4%)	8 (19.0%)	0.4352

Los datos se expresan en forma de mediana y rango intercuartil (IQR) para las variables continuas, y en número y porcentaje para las variables categóricas.

Tabla 16. Caracterización de los datos clínicos para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Grupo de riesgo clínico				
Grupo 0	31 (12.8%)	26 (12.9%)	5 (11.9%)	1.0000
Grupo 1	45 (18.5%)	38 (18.9%)	7 (16.7%)	0.8300
Grupo 2	60 (24.7%)	50 (24.9%)	10 (23.8%)	1.0000
Grupo 3	107 (44.0%)	87 (43.3%)	20 (47.6%)	0.6129
Anticoagulación				
No	201 (82.7%)	165 (82.1%)	36 (85.7%)	0.6596
Sí	42 (17.3%)	36 (17.9%)	6 (14.3%)	
Antiagregación				
No	187 (77.0%)	157 (78.1%)	30 (71.4%)	0.4199
Sí	56 (23.0%)	44 (21.9%)	12 (28.6%)	
Vacuna antigripal (año anterior)				
No	82 (33.7%)	71 (35.3%)	11 (26.2%)	0.2862
Sí	161 (66.3%)	130 (64.7%)	31 (73.8%)	
Vacuna antineumocócica				
No	151 (62.1%)	127 (63.2%)	24 (57.1%)	0.4874
Sí	92 (37.9%)	74 (36.8%)	18 (42.9%)	

Los datos se expresan en forma de número y porcentaje para las variables categóricas.

Respecto a las **comorbilidades** de los pacientes, las diferencias significativas entre grupos se dieron para las variables *Insuficiencia respiratoria* y *Taquiarritmia.* La insuficiencia respiratoria predominó en pacientes de clase positiva, mientras que la ausencia de la misma fue más frecuente en la clase negativa. Por otro lado, la taquiarritmia fue más común en los sujetos que reingresan que en los que no. Estos hallazgos pueden observarse en la Tabla 17.

Tabla 17 - I. Caracterización de las comorbilidades previas para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Hipertensión arterial				
No	88 (36.2%)	73 (36.3%)	15 (35.7%)	1.0000
Sí	155 (63.8%)	128 (63.7%)	27 (64.3%)	
Diabetes mellitus				
No	180 (74.1%)	148 (73.6%)	32 (76.2%)	0.8474
Sí	63 (25.9%)	53 (26.4%)	10 (23.8%)	
Dislipemia				
No	149 (61.3%)	122 (60.7%)	27 (64.3%)	0.7295
Sí	94 (38.7%)	79 (39.3%)	15 (35.7%)	
Cardiopatía isquémica				
No	225 (92.6%)	187 (93.0%)	38 (90.5%)	0.5249
Sí	18 (7.4%)	14 (7.0%)	4 (9.5%)	
Insuficiencia cardíaca				
No	218 (89.7%)	180 (89.6%)	38 (90.5%)	1.0000
Sí	25 (10.3%)	21 (10.4%)	4 (9.5%)	
Insuficiencia respiratoria				
No	122 (50.2%)	112 (55.7%)	10 (23.8%)	**0.0002**
Sí	121 (49.8%)	89 (44.3%)	32 (76.2%)	
Bronquiectasias				
No	194 (79.8%)	159 (79.1%)	35 (83.3%)	0.6736
Sí	49 (20.2%)	42 (20.9%)	7 (16.7%)	
Taquiarritmia				
No	230 (94.7%)	194 (96.5%)	36 (85.7%)	**0.0126**
Sí	13 (5.3%)	7 (3.5%)	6 (14.3%)	
Accidente cerebrovascular				
No	233 (95.9%)	192 (95.5%)	41 (97.6%)	1.0000
Sí	10 (4.1%)	9 (4.5%)	1 (2.4%)	
Neoplasia pulmonar				
No	232 (95.5%)	191 (95.0%)	41 (97.6%)	0.6950
Sí	11 (4.5%)	10 (5.0%)	1 (2.4%)	
Otras neoplasias				
No	190 (78.2%)	156 (77.6%)	34 (81.0%)	0.8373
Sí	53 (21.8%)	45 (22.4%)	8 (19.0%)	
Enfermedad renal				
No	239 (98.4%)	199 (99.0%)	40 (95.2%)	0.1392
Sí	4 (1.6%)	2 (1.0%)	2 (4.8%)	
Osteoporosis				
No	231 (95.1%)	190 (94.5%)	41 (97.6%)	0.6971
Sí	12 (4.9%)	11 (5.5%)	1 (2.4%)	

Los datos se expresan en forma de número y porcentaje para las variables categóricas.

Tabla 17 - II (cont.) Caracterización de las comorbilidades previas para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Ansiedad				
No	225 (92.6%)	185 (92.0%)	40 (95.2%)	0.7462
Sí	18 (7.4%)	16 (8.0%)	2 (4.8%)	
Depresión				
No	233 (95.9%)	194 (96.5%)	39 (92.9%)	0.3835
Sí	10 (4.1%)	7 (3.5%)	3 (7.1%)	
Anemia				
No	235 (96.7%)	195 (97.0%)	40 (95.2%)	0.6297
Sí	8 (3.3%)	6 (3.0%)	2 (4.8%)	
Tromboembolismo pulmonar				
No	232 (95.5%)	192 (95.5%)	40 (95.2%)	1.0000
Sí	11 (4.5%)	9 (4.5%)	2 (4.8%)	
Síndrome de apnea-hipopnea				
No	178 (73.3%)	146 (72.6%)	32 (76.2%)	0.7051
Sí	65 (26.7%)	55 (27.4%)	10 (23.8%)	

Los datos se expresan en forma de número y porcentaje para las variables categóricas.

En cuanto a la **espirometría previa al ingreso**, el grupo Reingreso mostró unos valores significativamente más bajos de *FVC basal (% teórico)* en comparación con los no reingresos. Asimismo, *FEV1 basal* y *FEV1 (% teórico)* presentaron valores significativamente inferiores en los pacientes que reingresan. Todo ello es recogido en la Tabla 18. Además, las Figuras 25-27 constituyen el diagrama de cajas de estas tres variables.

Tabla 18. Caracterización de la espirometría previa al ingreso para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
FVC basal	2.1 (1.7, 2.7)	2.2 (1.7, 2.8)	2.0 (1.5, 2.3)	0.0601
FVC basal (% teórico)	64.7 (54.7, 78.1)	66.0 (55.5, 79.3)	59.7 (52.0, 70.7)	**0.0375**
FVC basal (z score)	-2.1 (-2.9, -1.4)	-2.1 (-2.9, -1.3)	-2.4 (-3.1, -1.8)	0.0542
FEV1 basal	1.2 (0.8, 1.5)	1.2 (0.8, 1.6)	1.0 (0.8, 1.3)	**0.0186**
FEV1 basal (% teórico)	47.6 (35.4, 59.5)	48.9 (35.7, 61.0)	40.6 (32.8, 50.9)	**0.0164**
FEV1 basal (z score)	-2.9 (-3.5, -2.3)	-2.9 (-3.5, -2.2)	-3.2 (-3.9, -2.5)	0.0547
FEV1/FVC basal	0.6 (0.5, 0.6)	0.6 (0.5, 0.6)	0.5 (0.4, 0.6)	0.2549
FEV1/FVC basal (% teórico)	71.4 (60.5, 82.9)	72.1 (61.0, 83.0)	68.8 (55.8, 82.3)	0.3090
FEV1/FVC basal (z score)	-2.5 (-3.3, -1.5)	-2.5 (-3.3, -1.5)	-2.7 (-3.6, -1.5)	0.3889

Los datos se expresan en forma de mediana y rango intercuartil (IQR) para las variables continuas.

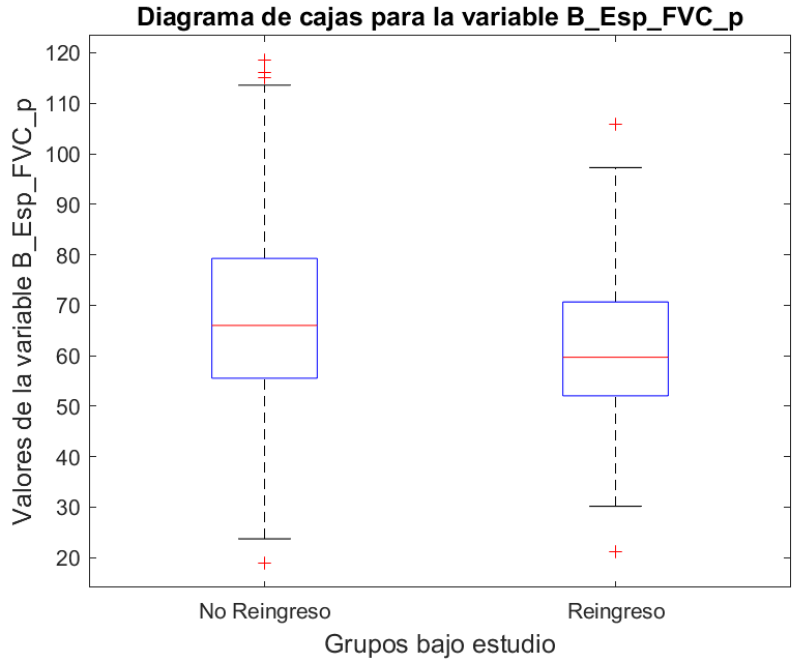

Figura 25. Diagrama de cajas para la variable "FVC basal (% teórico)". Denotación: *B_Esp_FVC_p.*

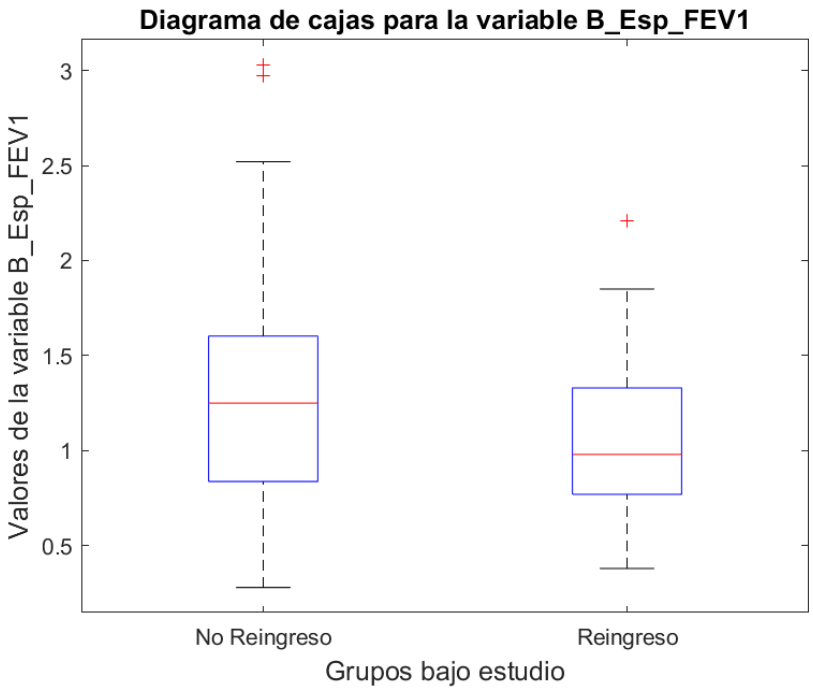

Figura 26. Diagrama de cajas para la variable "FEV1 basal". Denotación: *B_Esp_FEV1*.

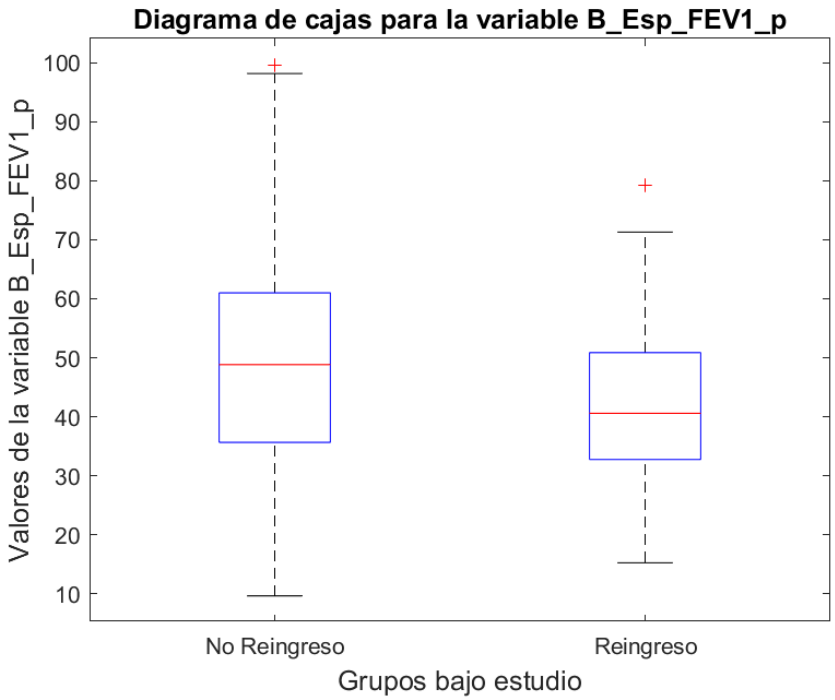

Figura 27. Diagrama de cajas para la variable "FEV1 basal (% teórico)". Denotación: *B_Esp_FEV1_p.*

En referencia a la **clasificación** de pacientes, se observaron diferencias significativas entre grupos para *Estratificación del riesgo, Fenotipo (GesEPOC), estadio GOLD* y *Número de ingresos por agudización (año previo).* En la estratificación del riesgo, fue significativamente más común el nivel de riesgo bajo para los pacientes que no reingresan, mientras que el riesgo alto fue predominante en los reingresos. En el fenotipo según GesEPOC, los sujetos no agudizadores predominaron en la clase negativa, mientras que los agudizadores con enfisema resultaron más habituales en la clase positiva. En el estadio GOLD, una proporción importante de los pacientes de la categoría B (menor gravedad) resultaron ser de la clase No Reingreso, mientras que la categoría D (mayor gravedad) fue significativamente más frecuente para los sujetos que reingresaron. Por otro lado, el número de ingresos por agudización en el año previo fue superior en el grupo positivo. Estos hallazgos se incluyen en la Tabla 19. Asimismo, la Figura 28 representa el diagrama de cajas de la variable *Número de ingresos por agudización (año previo).*

Tabla 19. Caracterización basal de la EPOC de acuerdo a las guías clínicas para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Estratificación del riesgo				
Bajo	51 (21.0%)	50 (24.9%)	1 (2.4%)	**0.0003**
Alto	192 (79.0%)	151 (75.1%)	41 (97.6%)	
Fenotipo (GesEPOC)				
No agudizador	105 (43.2%)	93 (46.3%)	12 (28.6%)	**0.0403**
Agudizador con enfisema	62 (25.5%)	44 (21.9%)	18 (42.9%)	**0.0065**
Agudizador con bronquitis crónica	33 (13.6%)	28 (13.9%)	5 (11.9%)	1.0000
Mixto	43 (17.7%)	36 (17.9%)	7 (16.7%)	1.0000
Grado obstrucción flujo aéreo (GOLD)				
Leve	14 (5.8%)	14 (7.0%)	0 (0.0%)	0.1373
Moderado	99 (40.7%)	87 (43.3%)	12 (28.6%)	0.0862
Grave	90 (37.0%)	70 (34.8%)	20 (47.6%)	0.1592
Muy grave	40 (16.5%)	30 (14.9%)	10 (23.8%)	0.1719
Estadio GOLD				
A	34 (14.0%)	32 (15.9%)	2 (4.8%)	0.0836
B	120 (49.4%)	108 (53.7%)	12 (28.6%)	0.0037
C	9 (3.7%)	8 (4.0%)	1 (2.4%)	1.0000
D	80 (32.9%)	53 (26.4%)	27 (64.3%)	**< 0.0001**
Número de ingresos por agudización (año previo)	0.0 (0.0, 1.0)	0.0 (0.0, 1.0)	1.5 (0.0, 3.0)	**< 0.0001**

Los datos se expresan en forma de número y porcentaje para las variables categóricas.

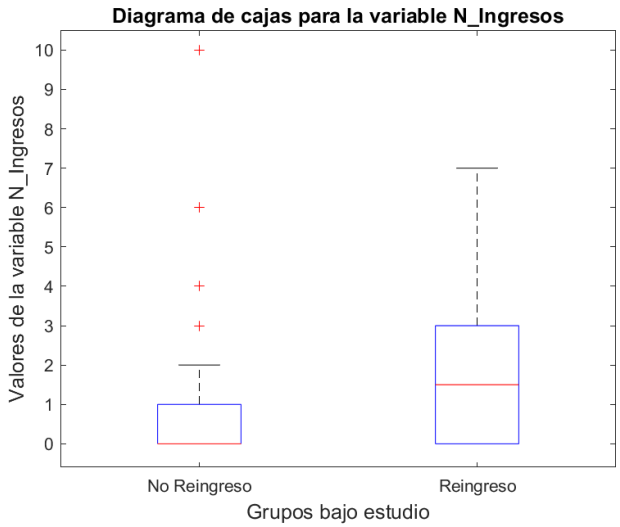

Figura 28. Diagrama de cajas para la variable "Número de ingresos por agudización (año previo)". Denotación: *N_Ingresos*.

En lo relativo a las variables relacionadas con la **terapia basal** recogidas en la Tabla 20**,** las características *Oxigenoterapia basal continua domiciliaria* y *Medicación de rescate basal: SABA* presentaron diferencias significativas entre clases. La pauta de oxigenoterapia domiciliaria fue más común en pacientes que reingresaron, mientras que la falta de esta fue más característica en sujetos que no reingresaron. Por otra parte, la administración de SABA fue más frecuente en pacientes de la clase positiva y la ausencia de su suministro fue más habitual en aquellos de la clase negativa.

Tabla 20 - I. Caracterización de la terapia basal para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Oxigenoterapia basal continua domiciliaria				
No	141 (58.0%)	126 (62.7%)	15 (35.7%)	**0.0018**
Sí	102 (42.0%)	75 (37.3%)	27 (64.3%)	
Ventilación no invasiva basal				
No	195 (80.2%)	165 (82.1%)	30 (71.4%)	0.1356
Sí	48 (19.8%)	36 (17.9%)	12 (28.6%)	
Medicación inhaladora basal: ninguna				
No	234 (96.3%)	192 (95.5%)	42 (100.0%)	0.3654
Sí	9 (3.7%)	9 (4.5%)	0 (0.0%)	
Medicación inhaladora basal: SABA				
No	229 (94.2%)	192 (95.5%)	37 (88.1%)	0.0726
Sí	14 (5.8%)	9 (4.5%)	5 (11.9%)	
Medicación inhaladora basal: SAMA				
No	229 (94.2%)	190 (94.5%)	39 (92.9%)	0.7147
Sí	14 (5.8%)	11 (5.5%)	3 (7.1%)	
Medicación inhaladora basal: LABA				
No	36 (14.8%)	32 (15.9%)	4 (9.5%)	0.3483
Sí	207 (85.2%)	169 (84.1%)	38 (90.5%)	
Medicación inhaladora basal: LAMA				
No	49 (20.2%)	44 (21.9%)	5 (11.9%)	0.2034
Sí	194 (79.8%)	157 (78.1%)	37 (88.1%)	
Medicación inhaladora basal: corticoides				
No	75 (30.9%)	67 (33.3%)	8 (19.0%)	0.0971
Sí	168 (69.1%)	134 (66.7%)	34 (81.0%)	

Los datos se expresan en forma de número y porcentaje para las variables categóricas.

Tabla 20 - II (cont.) Caracterización de la terapia basal para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Medicación de rescate basal: ninguna				
No	167 (68.7%)	136 (67.7%)	31 (73.8%)	0.4707
Sí	76 (31.3%)	65 (32.3%)	11 (26.2%)	
Medicación de rescate basal: SABA				
No	153 (63.0%)	133 (66.2%)	20 (47.6%)	**0.0341**
Sí	90 (37.0%)	68 (33.8%)	22 (52.4%)	
Medicación de rescate basal: SAMA				
No	230 (94.7%)	192 (95.5%)	38 (90.5%)	0.2481
Sí	13 (5.3%)	9 (4.5%)	4 (9.5%)	
Medicación de rescate basal: LABA				
No	235 (96.7%)	196 (97.5%)	39 (92.9%)	0.1434
Sí	8 (3.3%)	5 (2.5%)	3 (7.1%)	
Medicación de rescate basal: LAMA				
No	241 (99.2%)	200 (99.5%)	41 (97.6%)	0.3164
Sí	2 (0.8%)	1 (0.5%)	1 (2.4%)	
Medicación de rescate basal: corticoides				
No	233 (95.9%)	195 (97.0%)	38 (90.5%)	0.0738
Sí	10 (4.1%)	6 (3.0%)	4 (9.5%)	
Corticoterapia sistémica basal continua				
No	235 (96.7%)	194 (96.5%)	41 (97.6%)	1.0000
Sí	8 (3.3%)	7 (3.5%)	1 (2.4%)	
Teofilinas en estado basal				
No	217 (89.3%)	182 (90.5%)	35 (83.3%)	0.1748
Sí	26 (10.7%)	19 (9.5%)	7 (16.7%)	
IDFE4 en estado basal				
No	236 (97.1%)	195 (97.0%)	41 (97.6%)	1.0000
Sí	7 (2.9%)	6 (3.0%)	1 (2.4%)	
Mucolíticos en estado basal				
No	227 (93.4%)	188 (93.5%)	39 (92.9%)	0.7440
Sí	16 (6.6%)	13 (6.5%)	3 (7.1%)	
Antibióticos en estado basal				
No	241 (99.2%)	199 (99.0%)	42 (100.0%)	1.0000
Sí	2 (0.8%)	2 (1.0%)	0 (0.0%)	

Los datos se expresan en forma de número y porcentaje para las variables categóricas.

Tabla 20 - III (cont.) Caracterización de la terapia basal para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Rehabilitación respiratoria en estado basal				
No	239 (98.4%)	197 (98.0%)	42 (100.0%)	1.0000
Sí	4 (1.6%)	4 (2.0%)	0 (0.0%)	

Los datos se expresan en forma de número y porcentaje para las variables categóricas.

Respecto a los **test efectuados y sus puntuaciones**, las diferencias significativas se dieron en el test de Barthel, test de CAT y test EuroQol-5D (apartado de evaluación de movilidad, cuidado personal y actividades cotidianas), según muestra la Tabla 21. En el test de Barthel, los sujetos que reingresaron obtuvieron una puntuación inferior, pero superior en el test CAT que los no reingresos. En el EuroQoL-5D, una proporción baja de los pacientes que reingresaron reflejaron no tener ningún problema de movilidad. Los sujetos que constataron tener algunos problemas de movilidad fueron predominantemente los de clase positiva. Asimismo, los pacientes que reingresaron presentaron una mayor proporción con algunos problemas tanto en el cuidado personal como en la ejecución de actividades cotidianas.

En las Figuras 29 y 30, se representa el diagrama de cajas para el test de Barthel y de CAT, respectivamente.

Tabla 21 - I. Caracterización de los tests efectuados y sus puntuaciones para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Test_Barthel	95.0 (85.0, 100.0)	100.0 (85.0, 100.0)	90.0 (75.0, 100.0)	**0.0189**
Test_Charlson	5.0 (4.0, 6.0)	5.0 (4.0, 6.0)	5.0 (4.0, 7.0)	0.5648
Test de disnea (mMRC)				
0 puntos	16 (6.6%)	15 (7.5%)	1 (2.4%)	0.3193
1 punto	57 (23.5%)	50 (24.9%)	7 (16.7%)	0.3187
2 puntos	52 (21.4%)	44 (21.9%)	8 (19.0%)	0.8367
3 puntos	79 (32.5%)	63 (31.3%)	16 (38.1%)	0.4690
4 puntos	39 (16.0%)	29 (14.4%)	10 (23.8%)	0.1637
Test_CAT	23.0 (16.0, 28.0)	22.0 (15.8, 28.0)	27.0 (19.0, 32.0)	**0.0232**
Test TAI: adhesión				
Mala adhesión	30 (12.3%)	25 (12.4%)	5 (11.9%)	1.0000
Adhesión intermedia	43 (17.7%)	33 (16.4%)	10 (23.8%)	0.2688
Buena adhesión	170 (70.0%)	143 (71.1%)	27 (64.3%)	0.4592

Los datos se expresan en forma de mediana y rango intercuartil (IQR) para las variables continuas, y en número y porcentaje para las variables categóricas.

Tabla 21 - II (cont.) Caracterización de los tests efectuados y sus puntuaciones para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Test TAI: incumplimiento errático				
No	181 (74.5%)	152 (75.6%)	29 (69.0%)	0.4363
Sí	62 (25.5%)	49 (24.4%)	13 (31.0%)	
Test TAI: incumplimiento deliberado				
No	210 (86.4%)	173 (86.1%)	37 (88.1%)	1.0000
Sí	33 (13.6%)	28 (13.9%)	5 (11.9%)	
Test TAI: incumplimiento inconsciente				
No	53 (21.8%)	39 (19.4%)	14 (33.3%)	0.0631
Sí	190 (78.2%)	162 (80.6%)	28 (66.7%)	
Test EuroQoL-5D: movilidad				
Sin problemas	122 (50.2%)	108 (53.7%)	14 (33.3%)	**0.0179**
Algunos problemas	117 (48.1%)	89 (44.3%)	28 (66.7%)	**0.0105**
En cama	4 (1.6%)	4 (2.0%)	0 (0.0%)	1.0000
Test EuroQoL-5D: cuidado personal				
Sin problemas	170 (70.0%)	150 (74.6%)	20 (47.6%)	**0.0008**
Algunos problemas	62 (25.5%)	42 (20.9%)	20 (47.6%)	**0.0007**
Incapaz	11 (4.5%)	9 (4.5%)	2 (4.8%)	1.0000
Test EuroQoL-5D: actividades cotidianas				
Sin problemas	140 (57.6%)	123 (61.2%)	17 (40.5%)	**0.0163**
Algunos problemas	83 (34.2%)	61 (30.3%)	22 (52.4%)	**0.0076**
Incapaz	20 (8.2%)	17 (8.5%)	3 (7.1%)	1.0000
Test EuroQoL-5D: dolor/malestar				
Sin dolor	102 (42.0%)	89 (44.3%)	13 (31.0%)	0.1243
Dolor moderado	125 (51.4%)	99 (49.3%)	26 (61.9%)	0.1743
Mucho dolor	15 (6.2%)	12 (6.0%)	3 (7.1%)	0.7283
Test EuroQoL-5D: ansiedad/depresión				
No	134 (55.1%)	115 (57.2%)	19 (45.2%)	0.1745
Moderada	86 (35.4%)	68 (33.8%)	18 (42.9%)	0.2896
Muy ansioso/deprimido	23 (9.5%)	18 (9.0%)	5 (11.9%)	0.5636

Los datos se expresan en número y porcentaje para las variables categóricas.

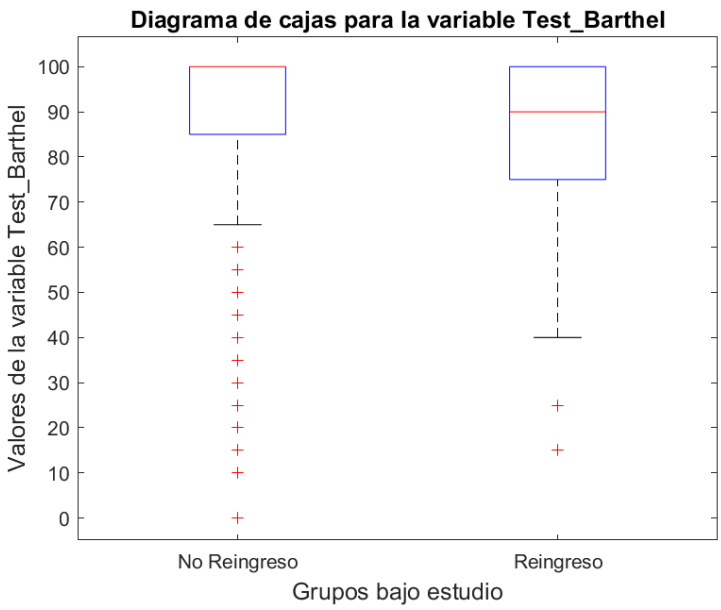

Figura 29. Diagrama de cajas para la variable "Test de Barthel". Denotación: *Test_Barthel*.

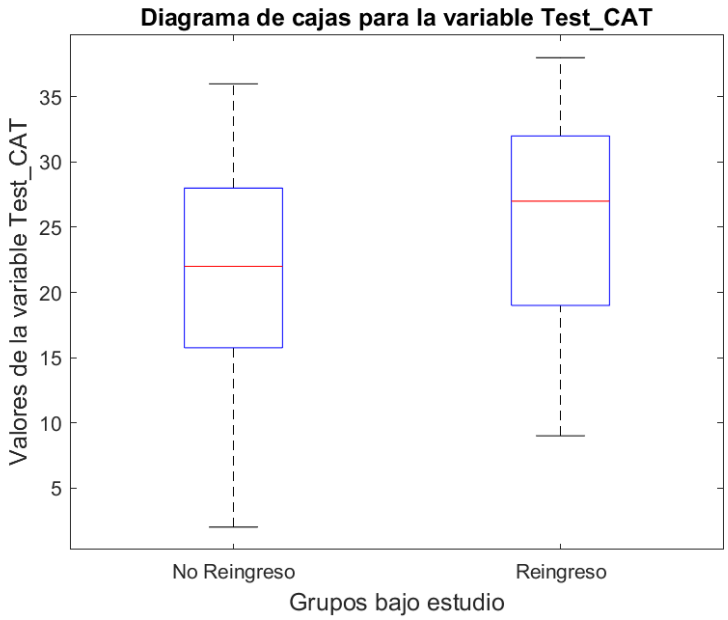

Figura 30. Diagrama de cajas para la variable "Test CAT". Denotación: *Test_CAT*.

En cuanto a las variables relacionadas con la **duración y motivo de ingreso** contenidas en la Tabla 22, las diferencias estadísticas significativas se muestran para la duración de la hospitalización, siendo superior en la clase positiva. Además, la Tabla 23 presenta las variables acerca de los **resultados de las pruebas realizadas durante** dicho **ingreso**, pudiéndose observar unos valores significativamente superiores en el grupo Reingreso en los niveles de PCO_2 y HCO_3.

Tabla 22. Caracterización de la duración y motivo de ingreso para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Número de días ingresado	7.0 (5.0, 10.0)	7.0 (5.0, 10.0)	8.5 (6.0, 14.0)	**0.0062**
Causa infecciosa				
No	102 (42.0%)	90 (44.8%)	12 (28.6%)	0.0597
Sí	141 (58.0%)	111 (55.2%)	30 (71.4%)	
Causa bacteriana				
No	183 (75.3%)	156 (77.6%)	27 (64.3%)	0.0781
Sí	60 (24.7%)	45 (22.4%)	15 (35.7%)	
Causa vírica				
No	195 (80.2%)	162 (80.6%)	33 (78.6%)	0.8314
Sí	48 (19.8%)	39 (19.4%)	9 (21.4%)	
Microorganismos resistentes				
No	233 (95.9%)	195 (97.0%)	38 (90.5%)	0.0738
Sí	10 (4.1%)	6 (3.0%)	4 (9.5%)	

Los datos se expresan en forma de mediana y rango intercuartil (IQR) para las variables continuas, y en número y porcentaje para las variables categóricas.

A continuación, se muestra el diagrama de cajas para la variable *Número de días ingresado* (Figura 31):

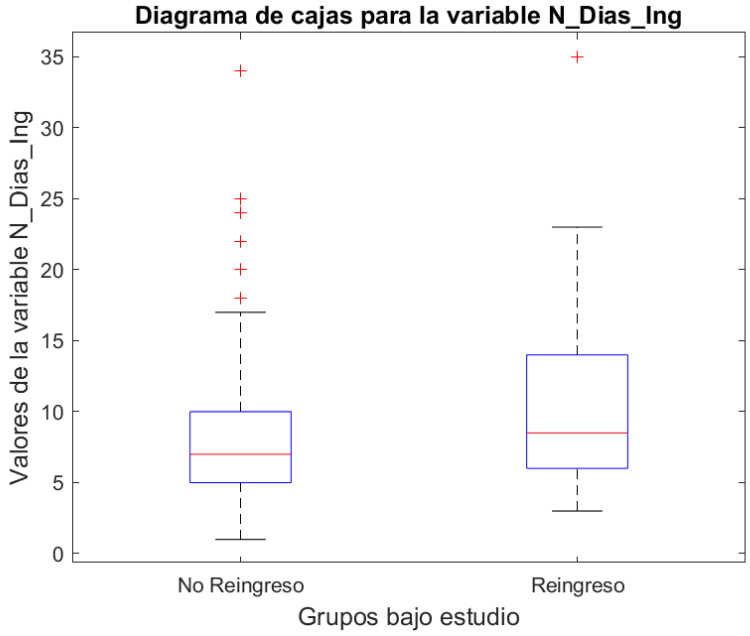

Figura 31. Diagrama de cajas para la variable "Número de días ingresado". Denotación: *N_Dias_Ing*.

Tabla 23. Caracterización de los resultados de pruebas realizadas al ingreso para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Tensión sistólica al ingreso	137.0 (122.0, 152.0)	137.0 (122.0, 151.0)	142.1 (125.0, 155.0)	0.2233
Tensión diastólica al ingreso	71.0 (62.0, 82.0)	72.0 (62.0, 82.0)	70.0 (61.0, 83.0)	0.6309
Frecuencia cardíaca al ingreso	96.0 (82.0, 107.0)	94.0 (81.8, 106.0)	102.5 (87.0, 114.0)	0.1919
pH al ingreso	7.4 (7.4, 7.5)	7.4 (7.4, 7.5)	7.4 (7.4, 7.4)	0.0798
PCO_2 al ingreso	43.0 (39.0, 50.0)	42.0 (38.0, 49.0)	47.5 (42.0, 53.0)	**0.0104**
PO_2 al ingreso	65.0 (54.0, 85.0)	64.0 (53.0, 89.0)	66.5 (57.0, 79.0)	0.5121
HCO_3 al ingreso	28.1 (24.9, 31.2)	27.9 (24.7, 30.6)	29.2 (26.0, 32.5)	**0.0336**
Leucocitos al ingreso	9900.0 (8000.0, 12900.0)	9800.0 (8000.0, 13000.0)	9900.0 (7500.0, 12100.0)	0.6284
Neutrófilos al ingreso	7812.2 (5627.6, 10361.2)	7838.3 (5747.5, 10370.7)	7710.8 (5616.0, 10331.1)	0.5929
Neutrófilos (%) al ingreso	78.2 (69.8, 85.3)	78.3 (69.5, 85.5)	76.8 (70.5, 83.2)	0.5406
Eosinófilos al ingreso	67.5 (16.4, 139.0)	68.0 (17.9, 142.0)	65.5 (12.9, 127.6)	0.5458
Eosinófilos (%) al ingreso	0.6 (0.1, 1.6)	0.6 (0.2, 1.7)	0.8 (0.1, 1.4)	0.7476

Los datos se expresan en forma de mediana y rango intercuartil (IQR) para las variables continuas.

Las Figuras 32 y 33 constituyen el diagrama de cajas para la variable *PCO_2 al ingreso* y *HCO_3 al ingreso,* respectivamente.

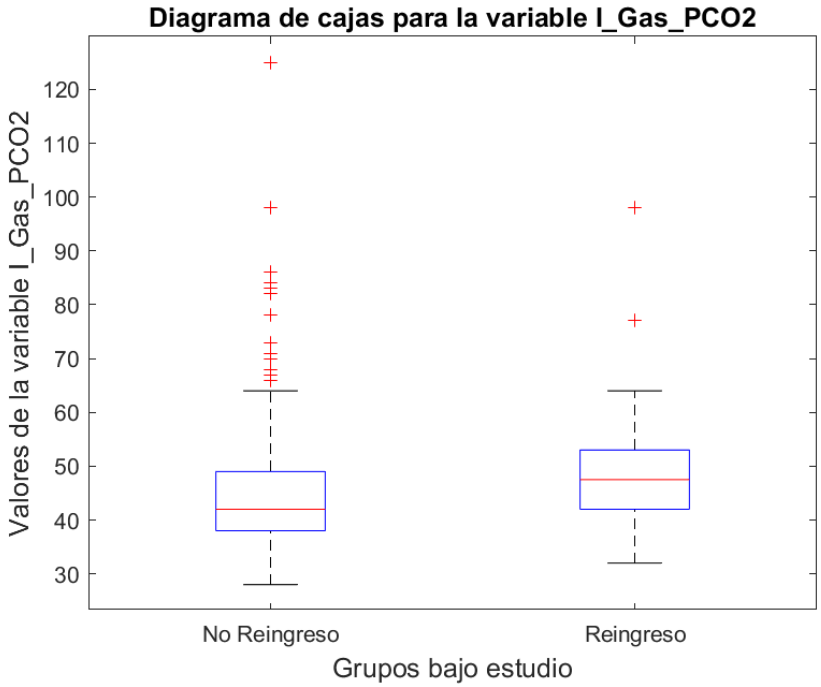

Figura 32. Diagrama de cajas para la variable "PCO_2 al ingreso". Denotación: *I_Gas_PCO2.*

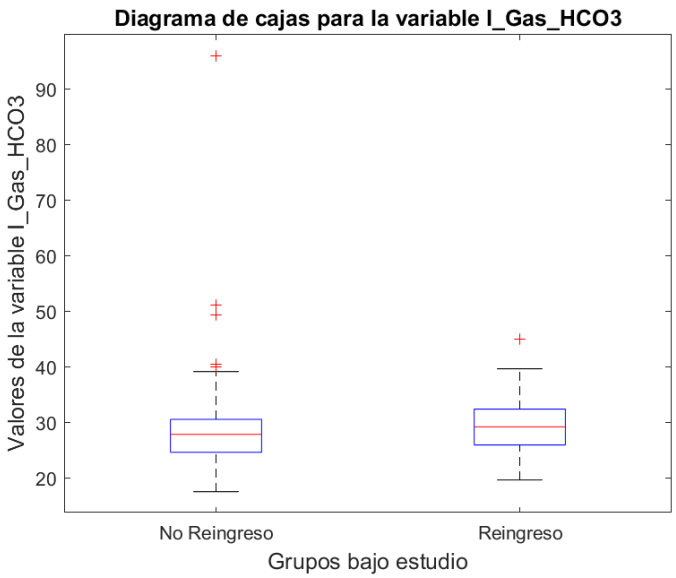

Figura 33. Diagrama de cajas para la variable "HCO_3 al ingreso". Denotación: *I_Gas_HCO3.*

Por otro lado, no se observaron diferencias significativas entre los grupos de estudio para las variables recogidas en la Tabla 24 relacionadas con los **síntomas y complicaciones al ingreso**.

Tabla 24 - I. Caracterización de los síntomas y complicaciones al ingreso para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Aumento de tos				
No	47 (19.3%)	36 (17.9%)	11 (26.2%)	0.2813
Sí	196 (80.7%)	165 (82.1%)	31 (73.8%)	
Aumento de disnea				
No	15 (6.2%)	13 (6.5%)	2 (4.8%)	1.0000
Sí	228 (93.8%)	188 (93.5%)	40 (95.2%)	
Aumento de expectoración				
No	94 (38.7%)	81 (40.3%)	13 (31.0%)	0.2985
Sí	149 (61.3%)	120 (59.7%)	29 (69.0%)	
Purulencia del esputo				
No	158 (65.0%)	134 (66.7%)	24 (57.1%)	0.2861
Sí	85 (35.0%)	67 (33.3%)	18 (42.9%)	
Dolor torácico				
No	203 (83.5%)	167 (83.1%)	36 (85.7%)	0.8206
Sí	40 (16.5%)	34 (16.9%)	6 (14.3%)	
Fiebre				
No	174 (71.6%)	144 (71.6%)	30 (71.4%)	1.0000
Sí	69 (28.4%)	57 (28.4%)	12 (28.6%)	

Los datos se expresan en forma de número y porcentaje para las variables categóricas.

Tabla 24 - II (cont.) Caracterización de los síntomas y complicaciones al ingreso para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Días de clínica que motivan el ingreso	4.0 (3.0, 7.0)	4.0 (3.0, 7.0)	4.1 (2.0, 7.0)	0.3150
Uso de musculatura accesoria				
No	156 (64.2%)	133 (66.2%)	23 (54.8%)	0.2149
Sí	87 (35.8%)	68 (33.8%)	19 (45.2%)	
Movimientos torácicos paradójicos				
No	225 (92.6%)	188 (93.5%)	37 (88.1%)	0.2089
Sí	18 (7.4%)	13 (6.5%)	5 (11.9%)	
Cianosis				
No	229 (94.2%)	189 (94.0%)	40 (95.2%)	1.0000
Sí	14 (5.8%)	12 (6.0%)	2 (4.8%)	
Edemas periféricos				
No	201 (82.7%)	165 (82.1%)	36 (85.7%)	0.6596
Sí	42 (17.3%)	36 (17.9%)	6 (14.3%)	
Inestabilidad hemodinámica				
No	238 (97.9%)	197 (98.0%)	41 (97.6%)	1.0000
Sí	5 (2.1%)	4 (2.0%)	1 (2.4%)	
Deterioro mental				
No	230 (94.7%)	189 (94.0%)	41 (97.6%)	0.7040
Sí	13 (5.3%)	12 (6.0%)	1 (2.4%)	
Disnea				
No	84 (34.6%)	72 (35.8%)	12 (28.6%)	0.4758
Sí	159 (65.4%)	129 (64.2%)	30 (71.4%)	
Arritmias				
No	234 (96.3%)	196 (97.5%)	38 (90.5%)	0.0506
Sí	9 (3.7%)	5 (2.5%)	4 (9.5%)	
Insuficiencia cardíaca				
No	217 (89.3%)	178 (88.6%)	39 (92.9%)	0.5850
Sí	26 (10.7%)	23 (11.4%)	3 (7.1%)	
Cardiopatía isquémica				
No	241 (99.2%)	200 (99.5%)	41 (97.6%)	0.3164
Sí	2 (0.8%)	1 (0.5%)	1 (2.4%)	
Derrame pleural				
No	237 (97.5%)	197 (98.0%)	40 (95.2%)	0.2775
Sí	6 (2.5%)	4 (2.0%)	2 (4.8%)	

Los datos se expresan en forma de mediana y rango intercuartil (IQR) para las variables continuas, y en número y porcentaje para las variables categóricas.

Tabla 24 – III (cont.) Caracterización de los síntomas y complicaciones al ingreso para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Neumonía				
No	209 (86.0%)	175 (87.1%)	34 (81.0%)	0.3282
Sí	34 (14.0%)	26 (12.9%)	8 (19.0%)	
Sepsis				
No	240 (98.8%)	199 (99.0%)	41 (97.6%)	0.4355
Sí	3 (1.2%)	2 (1.0%)	1 (2.4%)	
Insuficiencia respiratoria				
No	100 (41.2%)	81 (40.3%)	19 (45.2%)	0.6065
Sí	143 (58.8%)	120 (59.7%)	23 (54.8%)	

Los datos se expresan en forma de número y porcentaje para las variables categóricas.

En referencia a la **terapia al ingreso** (Tabla 25) **y al alta** (Tabla 26), se hallaron diferencias significativas entre clases en el uso de mucolíticos, la aplicación de ventilación no invasiva y la administración de teofilinas. La pauta de mucolíticos y ventilación no invasiva al ingreso resultaron más comunes en pacientes del grupo positivo (Reingreso). Las teofilinas, por su parte, fueron suministradas con una frecuencia significativamente superior en los pacientes que reingresaron.

Tabla 25 - I. Caracterización de la terapia al ingreso para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Medicación inhaladora al ingreso: ninguna				
No	241 (99.2%)	199 (99.0%)	42 (100.0%)	1.0000
Sí	2 (0.8%)	2 (1.0%)	0 (0.0%)	
Medicación inhaladora al ingreso: SABA				
No	54 (22.2%)	44 (21.9%)	10 (23.8%)	0.8387
Sí	189 (77.8%)	157 (78.1%)	32 (76.2%)	
Medicación inhaladora al ingreso: SAMA				
No	61 (25.1%)	50 (24.9%)	11 (26.2%)	0.8466
Sí	182 (74.9%)	151 (75.1%)	31 (73.8%)	

Los datos se expresan en forma de número y porcentaje para las variables categóricas.

Tabla 25 - II (cont.) Caracterización de la terapia al ingreso para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Medicación inhaladora al ingreso: LABA				
No	214 (88.1%)	179 (89.1%)	35 (83.3%)	0.2999
Sí	29 (11.9%)	22 (10.9%)	7 (16.7%)	
Medicación inhaladora al ingreso: LAMA				
No	185 (76.1%)	157 (78.1%)	28 (66.7%)	0.1166
Sí	58 (23.9%)	44 (21.9%)	14 (33.3%)	
Medicación inhaladora al ingreso: corticoides				
No	112 (46.1%)	98 (48.8%)	14 (33.3%)	0.0883
Sí	131 (53.9%)	103 (51.2%)	28 (66.7%)	
Corticoterapia sistémica al ingreso				
No	24 (9.9%)	23 (11.4%)	1 (2.4%)	0.0890
Sí	219 (90.1%)	178 (88.6%)	41 (97.6%)	
Ingreso en UVI				
No	237 (97.5%)	197 (98.0%)	40 (95.2%)	0.2775
Sí	6 (2.5%)	4 (2.0%)	2 (4.8%)	
Oxigenoterapia al ingreso				
No	23 (9.5%)	22 (10.9%)	1 (2.4%)	0.1419
Sí	220 (90.5%)	179 (89.1%)	41 (97.6%)	
Teofilinas en el ingreso				
No	224 (92.2%)	186 (92.5%)	38 (90.5%)	0.7508
Sí	19 (7.8%)	15 (7.5%)	4 (9.5%)	
IDFE4 en el ingreso				
No	238 (97.9%)	196 (97.5%)	42 (100.0%)	0.5908
Sí	5 (2.1%)	5 (2.5%)	0 (0.0%)	
Mucolíticos en el ingreso				
No	185 (76.1%)	160 (79.6%)	25 (59.5%)	**0.0090**
Sí	58 (23.9%)	41 (20.4%)	17 (40.5%)	
Ventilación no invasiva al ingreso				
No	200 (82.3%)	172 (85.6%)	28 (66.7%)	**0.0067**
Sí	43 (17.7%)	29 (14.4%)	14 (33.3%)	
Ventilación invasiva al ingreso				
No	241 (99.2%)	200 (99.5%)	41 (97.6%)	0.3164
Sí	2 (0.8%)	1 (0.5%)	1 (2.4%)	

Los datos se expresan en forma de número y porcentaje para las variables categóricas.

Tabla 25 - III (cont.) Caracterización de la terapia al ingreso para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Antibioterapia al ingreso				
No	21 (8.6%)	19 (9.5%)	2 (4.8%)	0.5446
Sí	222 (91.4%)	182 (90.5%)	40 (95.2%)	

Los datos se expresan en forma de número y porcentaje para las variables categóricas.

Tabla 26 - I. Caracterización de la terapia al alta para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Oxigenoterapia continua domiciliaria al alta				
No	109 (44.9%)	96 (47.8%)	13 (31.0%)	0.0601
Sí	134 (55.1%)	105 (52.2%)	29 (69.0%)	
Ventilación no invasiva al alta				
No	197 (81.1%)	166 (82.6%)	31 (73.8%)	0.1971
Sí	46 (18.9%)	35 (17.4%)	11 (26.2%)	
Medicación inhaladora al alta: ninguna				
No	239 (98.4%)	197 (98.0%)	42 (100.0%)	1.0000
Sí	4 (1.6%)	4 (2.0%)	0 (0.0%)	
Medicación inhaladora al alta: SABA				
No	224 (92.2%)	187 (93.0%)	37 (88.1%)	0.3385
Sí	19 (7.8%)	14 (7.0%)	5 (11.9%)	
Medicación inhaladora al alta: SAMA				
No	224 (92.2%)	188 (93.5%)	36 (85.7%)	0.1097
Sí	19 (7.8%)	13 (6.5%)	6 (14.3%)	
Medicación inhaladora al alta: LABA				
No	39 (16.0%)	33 (16.4%)	6 (14.3%)	0.8214
Sí	204 (84.0%)	168 (83.6%)	36 (85.7%)	
Medicación inhaladora al alta: LAMA				
No	40 (16.5%)	33 (16.4%)	7 (16.7%)	1.0000
Sí	203 (83.5%)	168 (83.6%)	35 (83.3%)	
Medicación inhaladora al alta: corticoides				
No	67 (27.6%)	58 (28.9%)	9 (21.4%)	0.4476
Sí	176 (72.4%)	143 (71.1%)	33 (78.6%)	
Medicación de rescate al alta: ninguna				
No	174 (71.6%)	144 (71.6%)	30 (71.4%)	1.0000
Sí	69 (28.4%)	57 (28.4%)	12 (28.6%)	

Los datos se expresan en forma de número y porcentaje para las variables categóricas.

Tabla 26 - II (cont.) Caracterización de la terapia al alta para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Medicación de rescate al alta: SABA				
No	156 (64.2%)	135 (67.2%)	21 (50.0%)	0.0505
Sí	87 (35.8%)	66 (32.8%)	21 (50.0%)	
Medicación de rescate al alta: SAMA				
No	234 (96.3%)	195 (97.0%)	39 (92.9%)	0.1898
Sí	9 (3.7%)	6 (3.0%)	3 (7.1%)	
Medicación de rescate al alta: LABA				
No	234 (96.3%)	194 (96.5%)	40 (95.2%)	0.6561
Sí	9 (3.7%)	7 (3.5%)	2 (4.8%)	
Medicación de rescate al alta: LAMA				
No	238 (97.9%)	196 (97.5%)	42 (100.0%)	0.5908
Sí	5 (2.1%)	5 (2.5%)	0 (0.0%)	
Medicación de rescate al alta: corticoides				
No	229 (94.2%)	190 (94.5%)	39 (92.9%)	0.7147
Sí	14 (5.8%)	11 (5.5%)	3 (7.1%)	
Corticoterapia sistémica continua al alta				
No	93 (38.3%)	80 (39.8%)	13 (31.0%)	0.3015
Sí	150 (61.7%)	121 (60.2%)	29 (69.0%)	
Antibióticos al alta				
No	118 (48.6%)	95 (47.3%)	23 (54.8%)	0.4003
Sí	125 (51.4%)	106 (52.7%)	19 (45.2%)	
Teofilinas al alta				
No	210 (86.4%)	178 (88.6%)	32 (76.2%)	**0.0459**
Sí	33 (13.6%)	23 (11.4%)	10 (23.8%)	
IDFE4 al alta				
No	237 (97.5%)	196 (97.5%)	41 (97.6%)	1.0000
Sí	6 (2.5%)	5 (2.5%)	1 (2.4%)	
Mucolíticos al alta				
No	194 (79.8%)	165 (82.1%)	29 (69.0%)	0.0884
Sí	49 (20.2%)	36 (17.9%)	13 (31.0%)	
Rehabilitación respiratoria al alta				
No	228 (93.8%)	187 (93.0%)	41 (97.6%)	0.4792
Sí	15 (6.2%)	14 (7.0%)	1 (2.4%)	

Los datos se expresan en forma de número y porcentaje para las variables categóricas.

Por último, en lo relativo a los **datos relacionados con el reingreso** incluidos en la Tabla 27, se muestran diferencias significativas entre grupos en la cantidad de visitas a urgencias en los 30 días posteriores al alta. La clase prioritaria fue la positiva, es decir, acudieron más a urgencias los pacientes que finalmente reingresaron que los que no.

Tabla 27. Caracterización de los datos relacionados con el reingreso para las dos clases bajo estudio en la cohorte retrospectiva.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)	p-valor
Días desde el alta al exitus	830.0 (421.0, 1120.0)	880.0 (437.2, 1117.0)	662.0 (178.5, 1180.8)	0.5063
Exitus durante los 30 días post-alta				
No	238 (97.9%)	198 (98.5%)	40 (95.2%)	0.2070
Sí	5 (2.1%)	3 (1.5%)	2 (4.8%)	
Visitas a Urgencias (en los 30 días post-alta)				
0 visitas	193 (79.4%)	189 (94.0%)	4 (9.5%)	**< 0.0001**
1 visita	45 (18.5%)	11 (5.5%)	34 (81.0%)	**< 0.0001**
2 visitas	4 (1.6%)	1 (0.5%)	3 (7.1%)	**0.0171**
3 visitas	1 (0.4%)	0 (0.0%)	1 (2.4%)	0.1728
Visitas a consultas de Neumología (en los 30 días post-alta)				
0 visitas	206 (85.1%)	166 (83.0%)	40 (95.2%)	0.0540
1 visita	35 (14.5%)	33 (16.5%)	2 (4.8%)	0.0539
2 visitas	1 (0.4%)	1 (0.5%)	0 (0.0%)	1.0000

Los datos se expresan en forma de mediana y rango intercuartil (IQR) para las variables continuas, y en número y porcentaje para las variables categóricas.

5.3.2. Base de datos prospectiva

En la Tabla 28 se resume la caracterización de las variables que conforman la base de datos prospectiva. En ella, se observan ciertas diferencias entre el paciente que reingresó (clase 1) y los pacientes que no reingresaron (clase 0). El sujeto del grupo positivo muestra una mediana del número de ingresos por agudización en el año previo superior a los sujetos del grupo negativo, así como un mayor empleo de mucolíticos al ingreso y en estado basal, oxigenoterapia basal continua domiciliaria e incumplimiento errático según el test TAI. No obstante, el limitado tamaño muestral para la clase positiva impide generalizar.

Respecto a los microorganismos resistentes, movimientos torácicos paradójicos, anemia, neumonía y ventilación no invasiva al ingreso, no se registró ningún paciente con su presencia.

Tabla 28 - I. Caracterización de las variables de la cohorte prospectiva para las dos clases bajo estudio.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)
Número de ingresos por agudización (año previo)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	3.0 (3.0, 3.0)
Mucolíticos en el ingreso			
No	6 (60.0%)	6 (66.7%)	0 (0.0%)
Sí	4 (40.0%)	3 (33.3%)	1 (100.0%)
Mucolíticos al alta			
No	9 (90.0%)	8 (88.9%)	1 (100.0%)
Sí	1 (10.0%)	1 (11.1%)	0 (0.0%)
Microorganismos resistentes			
No	10 (100.0%)	9 (100.0%)	1 (100.0%)
Sí	0 (0.0%)	0 (0.0%)	0 (0.0%)
Test TAI: incumplimiento inconsciente			
No	6 (60.0%)	5 (55.6%)	1 (100.0%)
Sí	4 (40.0%)	4 (44.4%)	0 (0.0%)
Oxigenoterapia basal continua domiciliaria			
No	6 (60.0%)	6 (66.7%)	0 (0.0%)
Sí	4 (40.0%)	3 (33.3%)	1 (100.0%)
Teofilinas al alta			
No	9 (90.0%)	8 (88.9%)	1 (100.0%)
Sí	1 (10.0%)	1 (11.1%)	0 (0.0%)
Uso de musculatura accesoria			
No	8 (80.0%)	7 (77.8%)	1 (100.0%)
Sí	2 (20.0%)	2 (22.2%)	0 (0.0%)
Causa bacteriana			
No	6 (60.0%)	5 (55.6%)	1 (100.0%)
Sí	4 (40.0%)	4 (44.4%)	0 (0.0%)
Teofilinas en estado basal			
No	9 (90.0%)	8 (88.9%)	1 (100.0%)
Sí	1 (10.0%)	1 (11.1%)	0 (0.0%)
Grado obstrucción flujo aéreo (GOLD)			
Leve	0 (0.0%)	0 (0.0%)	0 (0.0%)
Moderado	2 (20.0%)	2 (22.2%)	0 (0.0%)
Grave	5 (50.0%)	4 (44.4%)	1 (100.0%)
Muy grave	3 (30.0%)	3 (33.3%)	0 (0.0%)
Medicación inhaladora al ingreso: LABA			
No	7 (70.0%)	6 (66.7%)	1 (100.0%)
Sí	3 (30.0%)	3 (33.3%)	0 (0.0%)

Los datos se expresan en forma de mediana y rango intercuartil (IQR) para las variables continuas, y en número y porcentaje para las variables categóricas.

Tabla 28 - II (cont.) Caracterización de las variables de la cohorte prospectiva para las dos clases bajo estudio.

Variable	Total sujetos	Clase 0 (no reingreso)	Clase 1 (reingreso)
Medicación inhaladora al ingreso: SABA			
No	4 (40.0%)	4 (44.4%)	0 (0.0%)
Sí	6 (60.0%)	5 (55.6%)	1 (100.0%)
Test EuroQoL-5D: cuidado personal			
Sin problemas	6 (60.0%)	5 (55.6%)	1 (100.0%)
Algunos problemas	4 (40.0%)	4 (44.4%)	0 (0.0%)
Incapaz	0 (0.0%)	0 (0.0%)	0 (0.0%)
Movimientos torácicos paradójicos			
No	10 (100.0%)	9 (100.0%)	1 (100.0%)
Sí	0 (0.0%)	0 (0.0%)	0 (0.0%)
Arritmias			
No	9 (90.0%)	8 (88.9%)	1 (100.0%)
Sí	1 (10.0%)	1 (11.1%)	0 (0.0%)
Teofilinas en el ingreso			
No	8 (80.0%)	7 (77.8%)	1 (100.0%)
Sí	2 (20.0%)	2 (22.2%)	0 (0.0%)
Anemia			
No	10 (100.0%)	9 (100.0%)	1 (100.0%)
Sí	0 (0.0%)	0 (0.0%)	0 (0.0%)
Medicación inhaladora al alta: SABA			
No	8 (80.0%)	7 (77.8%)	1 (100.0%)
Sí	2 (20.0%)	2 (22.2%)	0 (0.0%)
Edemas periféricos			
No	9 (90.0%)	9 (100.0%)	0 (0.0%)
Sí	1 (10.0%)	0 (0.0%)	1 (100.0%)
Ventilación no invasiva al ingreso			
No	10 (100.0%)	9 (100.0%)	1 (100.0%)
Sí	0 (0.0%)	0 (0.0%)	0 (0.0%)
Mucolíticos en estado basal			
No	8 (80.0%)	8 (88.9%)	0 (0.0%)
Sí	2 (20.0%)	1 (11.1%)	1 (100.0%)
Test TAI: incumplimiento errático			
No	6 (60.0%)	6 (66.7%)	0 (0.0%)
Sí	4 (40.0%)	3 (33.3%)	1 (100.0%)
Neumonía			
No	10 (100.0%)	9 (100.0%)	1 (100.0%)
Sí	0 (0.0%)	0 (0.0%)	0 (0.0%)

Los datos se expresan en forma de número y porcentaje para las variables categóricas.

5.4. Selección de variables

El algoritmo *ReliefF,* explicado en la sección de metodología, seleccionó las variables mostradas en la Tabla 29 para distintos valores de *K.* Se adjuntan también las puntuaciones o pesos de cada una de las características. Se observó que, independientemente del valor de *K* aplicado, algunas de las variables más relevantes, aunque los pesos asignados a ellas difiriesen.

Como se indicó previamente, finalmente se utilizó el subconjunto de variables obtenido para *K =5.* No obstante, la tabla incluye resultados con otros valores para evidenciar la coincidencia mencionada.

Tabla 29 - I. Variables seleccionadas con *ReliefF* para distintos valores de *K*.

***K* = 5**		***K* = 10**		***K* = 15**	
Variable	**Peso**	**Variable**	**Peso**	**Variable**	**Peso**
Mucolíticos al ingreso	0.158	**Microorganismos resistentes**	0.1347	**Microorganismos resistentes**	0.1255
Mucolíticos al alta	0.1461	**Mucolíticos al alta**	0.1112	**Teofilinas al alta**	0.0977
Microorganismos resistentes	0.1366	**Mucolíticos al ingreso**	0.1082	**Mucolíticos al ingreso**	0.0959
Test_TAI_I_Inc	0.1265	**Causa bacteriana**	0.0986	**Mucolíticos al alta**	0.0956
Oxigenoterapia domiciliaria basal	0.1176	**Uso musculatura accesoria**	0.0948	**Oxigenoterapia domiciliaria basal**	0.0923
Uso musculatura accesoria	0.0955	**Gold**	0.0907	**Gold**	0.0916
Teofilinas al alta	0.0903	**Oxigenoterapia domiciliaria basal**	0.0899	**I_Min_LABA**	0.0891
Causa bacteriana	0.0900	**Teofilinas al alta**	0.0869	**Teofilinas en estado basal**	0.0864
Gold	0.0872	**Teofilinas en estado basal**	0.0864	**Causa bacteriana**	0.0837
Teofilinas en estado basal	0.0852	**I_Min_LABA**	0.0838	**I_Min_SABA**	0.0806
I_Min_LABA	0.0833	**Test_TAI_I_Inc**	0.0802	**Uso musculatura accesoria**	0.0782
Test_E5D_CPe	0.0791	**Complicación: arritmias**	0.0731	**Test_TAI_I_Inc**	0.0723
Movimientos torácicos	0.0771	**Test_E5D_CPe**	0.0716	**Complicación: arritmias**	0.0713
I_Min_SABA	0.0742	**I_Min_SABA**	0.0711	**I_VNI**	0.0641
Complicación: arritmias	0.0704	**N_Ingresos**	0.0677	**N_Ingresos**	0.0636
N_Ingresos	0.0642	**Movimientos torácicos**	0.0623	**Movimientos torácicos**	0.0622
Teofilinas al ingreso	0.0598	**I_VNI**	0.0613	**Test_E5D_CPe**	0.0620

Test_TAI_I_Inc: test de TAI, incumplimiento inconsciente de la pauta del inhalador; Test_E5D_CPe: test EuroQoL-5D, cuidado personal; I_Min_LABA: medicación inhaladora al ingreso, LABA; I_Min_SABA: medicación inhaladora al ingreso, SABA; N_Ingresos: número de ingresos por agudización (año previo); I_VNI: ventilación no invasiva al ingreso.

Tabla 29 - II (cont.). Variables seleccionadas con *ReliefF* para distintos valores de *K*.

K = 5		*K* = 10		*K* = 15	
Variable	**Peso**	**Variable**	**Peso**	**Variable**	**Peso**
A_Min_SABA	0.0547	**Anemia**	0.0517	**Hogar**	0.0553
Anemia	0.0514	**A_Mre_SABA**	0.0486	**Taquiarritmia**	0.0549
Edemas periféricos	0.0468	**Aumento de tos**	0.0469	**I_Min_LAMA**	0.0544
I_VNI	0.0451	**B_Mre_SABA**	0.0468	**Teofilinas al ingreso**	0.0532
Test_TAI_I_Err	0.0443	**A_Min_SABA**	0.0455	**B_Mre_CI**	0.0499
Complicación: neumonía	0.0436	**Test_TAI_I_Err**	0.0453	**Complicación: neumonía**	0.0454
Mucolíticos en estado basal	0.0424	**Hogar**	0.0436	**B_Mre_SABA**	0.0432

A_Min_SABA: medicación inhaladora al alta, SABA; Test_TAI_I_Err: test de TAI, incumplimiento errático de la pauta del inhalador; Test_E5D_CPe: test EuroQoL-5D, cuidado personal; A_Mre_SABA: medicación de rescate al alta, SABA; B_Mre_SABA: mediación de rescate basal, SABA; I_Min_LAMA: medicación inhaladora al ingreso, LAMA; B_Mre_CI: medicación de rescate basal, corticoides; I_VNI: ventilación no invasiva al ingreso.

Para una mayor claridad visual, se adjunta a continuación un diagrama de barras de las variables seleccionadas con *K = 5* y sus correspondientes pesos (Figura 34):

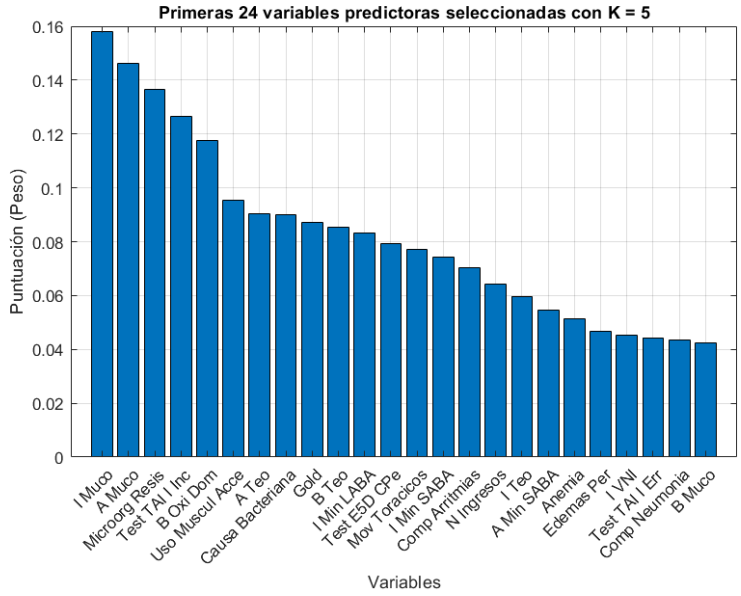

Figura 34. Diagrama de barras de las 24 variables de mayor relevancia según *ReliefF* y sus correspondientes pesos para *K = 5.*

I Muco: mucolíticos en el ingreso; A Muco: mucolíticos al alta; Microorg Resis: microorganismos resistentes; Test TAI I Inc: test de TAI, incumplimiento inconsciente de la pauta del inhalador; B Oxi Dom: oxigenoterapia continua domiciliaria basal; Uso Muscul Acce: uso de musculatura accesoria; A Teo: teofilinas al alta; B Teo: teofilinas en estado basal; I Min LABA: medicación inhaladora al ingreso, LABA; Test E5D CPe: test EuroQoL-5D, cuidado personal; Mov Torácicos: movimientos torácicos paradójicos; I Min SABA: medicación inhaladora al ingreso, SABA; Comp Arritmias: complicación de arritmias; N Ingresos: número de ingresos por agudización (año previo); I Teo: teofilinas en el ingreso; A Min SABA: medicación inhaladora al alta, SABA; Edemas Per: edemas periféricos; I VNI: ventilación no invasiva al ingreso; Test TAI I Err: test de TAI, incumplimiento errático de la pauta del inhalador; Comp Neumonía: complicación de neumonía; B Muco: mucolíticos en estado basal.

5.5. Diseño de modelos predictivos y optimización de sus hiperparámetros

Tras determinar las variables seleccionadas, estas sirvieron como entradas para los dos modelos predictivos desarrollados. En este apartado, se exponen los resultados de la optimización de hiperparámetros.

5.5.1. *Random Forest*

En la Tabla 30, se recogen los hiperparámetros finalmente aplicados tras la búsqueda orientada a maximizar la métrica de rendimiento *F1- score* en el conjunto de validación. Asimismo, en las Figuras 35-40 se muestra la evolución de la métrica *F1 score* en función del rango de valores considerado para cada hiperparámetro dentro del procedimiento de optimización secuencial descrito en el capítulo de metodología. El máximo de cada una de las gráficas se muestra marcado con un círculo rojo, indicando el punto óptimo finalmente elegido y aplicado en el entrenamiento final del modelo.

Tabla 30. Valores óptimos de los hiperparámetros para el modelo basado en *Random Forest* sobre la base de datos retrospectiva.

Hiperparámetro	Valor optimizado
Número de árboles	215
Penalización para los falsos positivos	2.5
Penalización para los falsos negativos	23
Tamaño mínimo de hoja	15
Número de predictores a muestrear	4
Número máximo de divisiones	80
Umbral de predicción	0.7

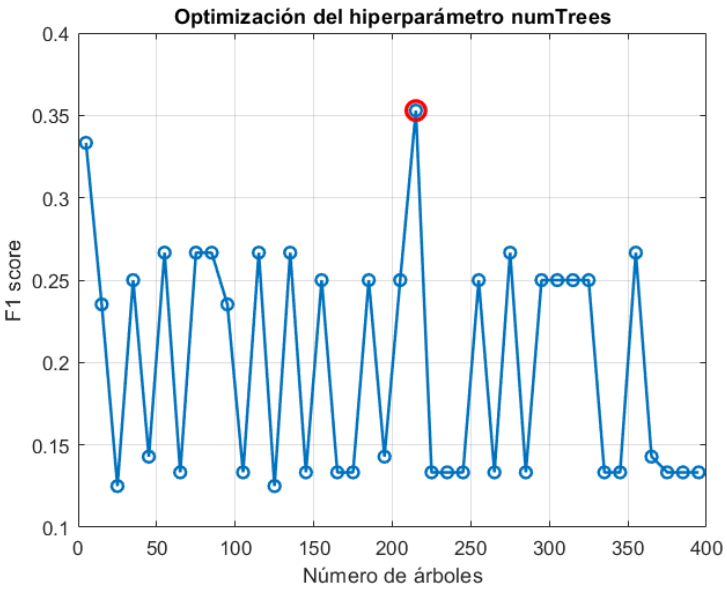

Figura 35. Optimización del número de árboles (*numTrees*) según *F1 score*.

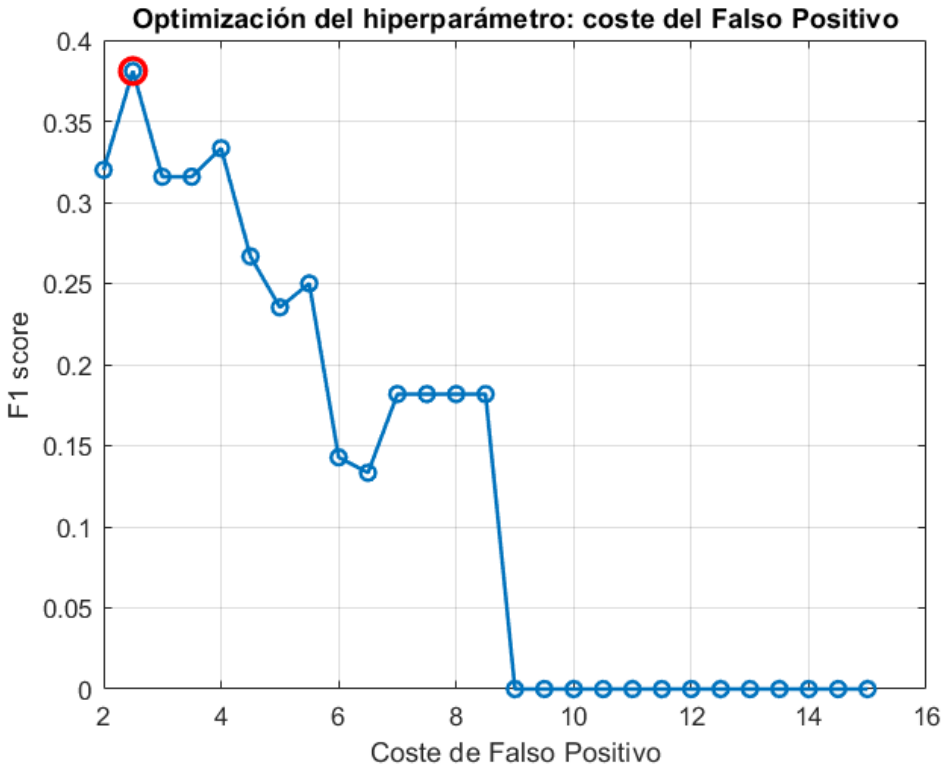

Figura 36. Optimización de la penalización de los falsos positivos según *F1 score*.

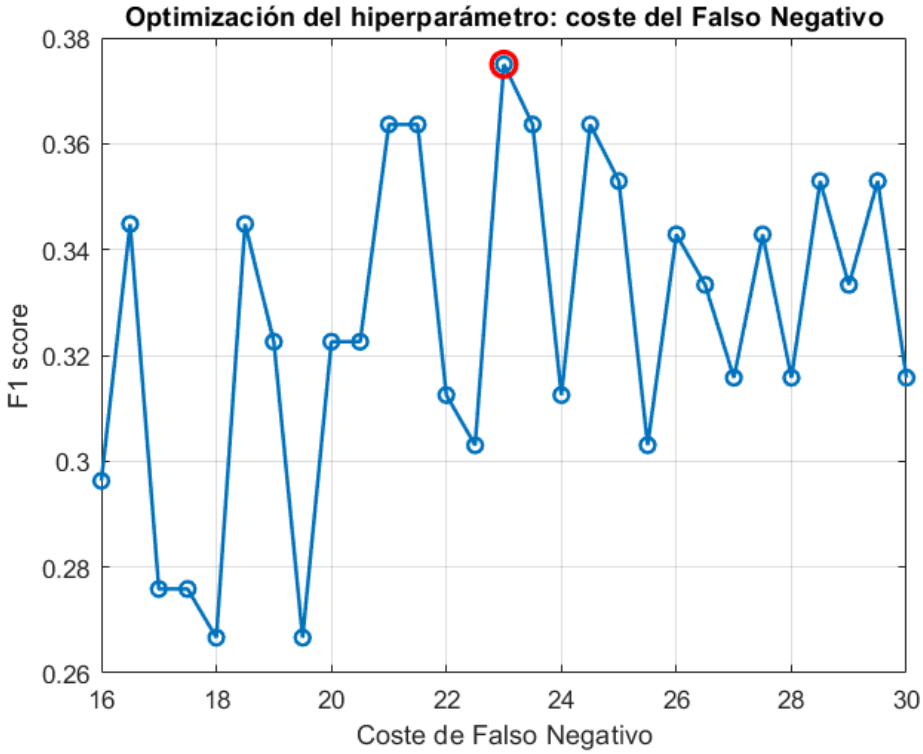

Figura 37. Optimización de la penalización de los falsos negativos según *F1 score*.

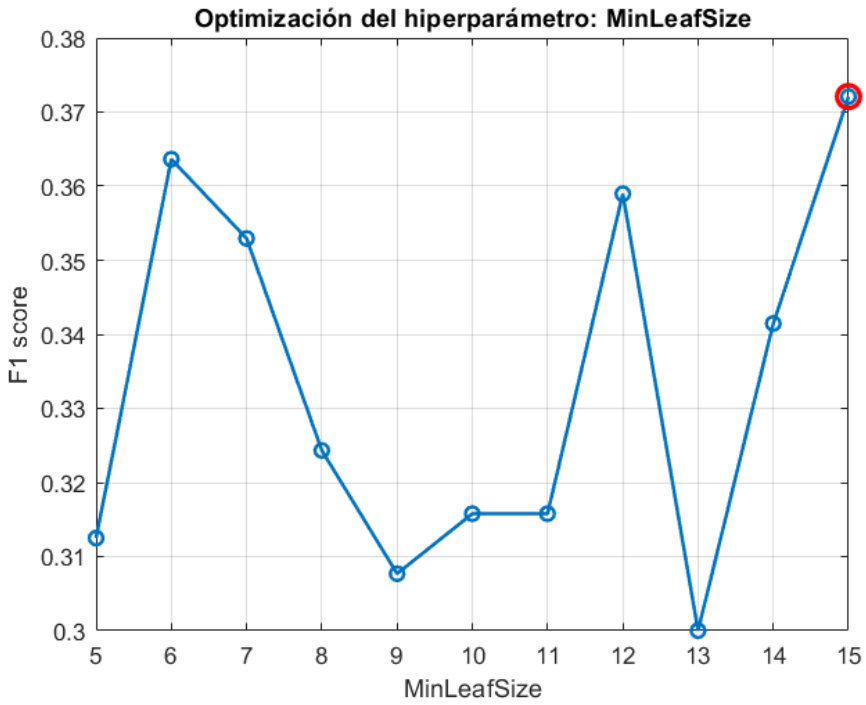

Figura 38. Optimización del tamaño mínimo de hoja (*MinLeafSize*) según *F1 score*.

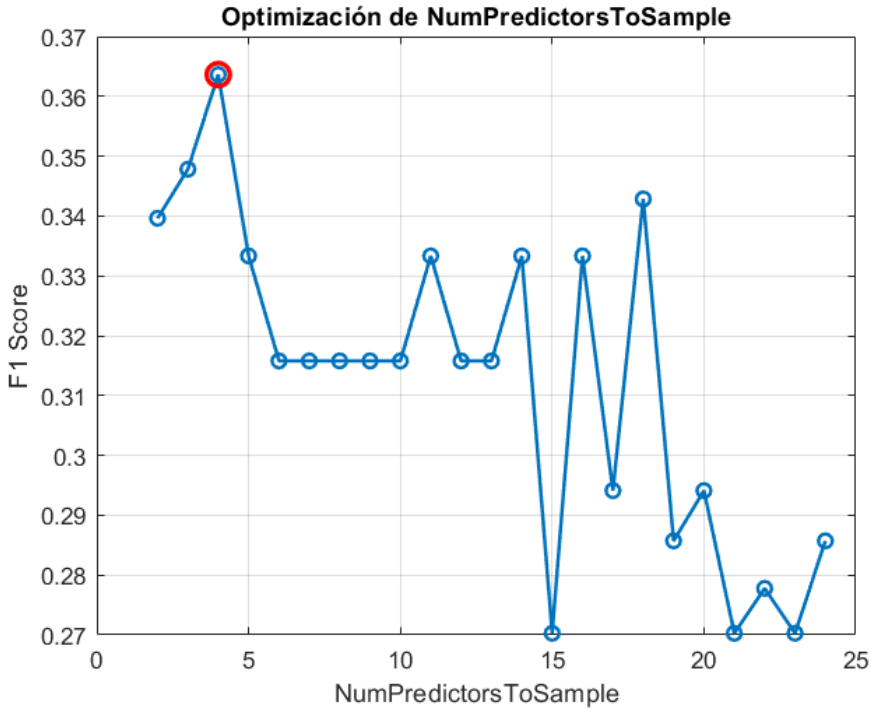

Figura 39. Optimización del número de predictores a muestrear (*NumPredictorsToSample*) según *F1 score.*

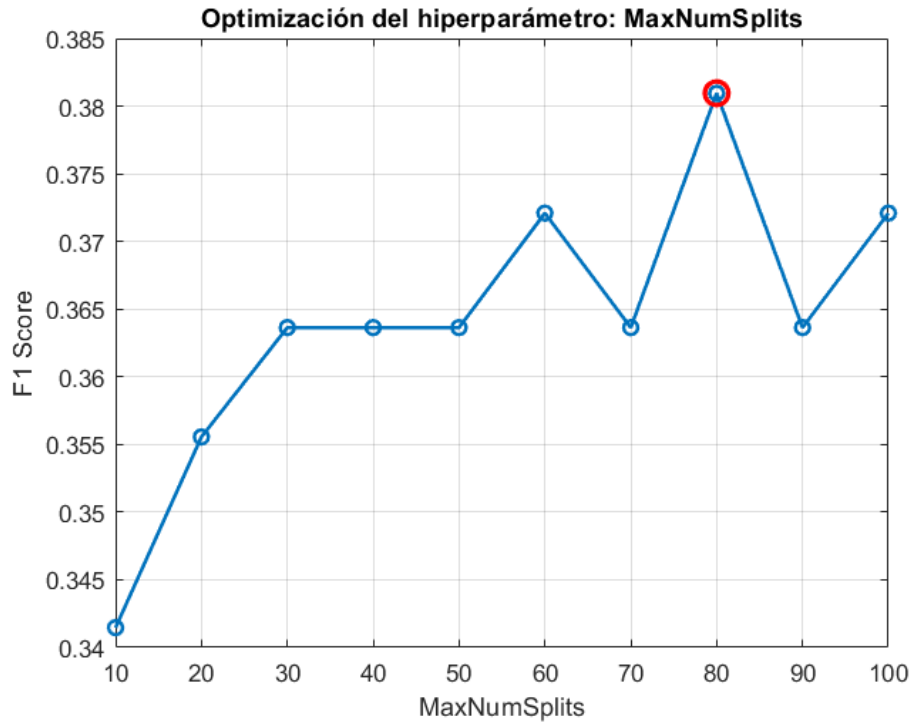

Figura 40. Optimización del máximo número de divisiones (*MaxNumSplits*) según *F1 score*.

5.5.2. Red neuronal perceptrón multicapa (MLP)

Los hiperparámetros a optimizar fueron el número de neuronas en la capa oculta, el parámetro de regularización y el umbral de decisión. Como se explicó en la sección de metodología de este trabajo, la optimización de cada uno de ellos se efectuó estableciendo el valor que maximizase la métrica de *F1 score* en el conjunto de validez. Es decir, se mantuvo el mismo criterio que en el modelo basado en *Random Forest*.

La Tabla 31 expone los valores finalmente asignados para estos hiperparámetros:

Tabla 31. Valores óptimos de los hiperparámetros para el modelo basado en un MLP sobre la base de datos retrospectiva.

Hiperparámetro	Valor optimizado
Neuronas en la capa oculta	28
Parámetro de regularización	0.2
Umbral de decisión	0.3

La Figura 41 exhibe el *F1 score* resultante de las distintas combinaciones de los hiperparámetros del número de neuronas en la capa oculta y el parámetro de regularización denotado como *alpha*. Con ello, se pretende mostrar de manera visual, el punto máximo obtenido para dicha métrica que representa la configuración óptima de la red. En este caso, se han identificado dos combinaciones óptimas, cuyo valor de *F1 score* es del 52.6%:

1 Número de neuronas en capa oculta: **28**. Parámetro de regularización (*alpha*): **0.2.**

2 Número de neuronas en capa oculta: **38**. Parámetro de regularización (*alpha*): **0.25.**

Puesto que ambas combinaciones ofrecen el mismo rendimiento, se optó por la configuración de 28 neuronas y *alpha* = 0.2, evitando el desarrollo de una red neuronal de mayor complejidad innecesariamente y promoviendo su eficiencia.

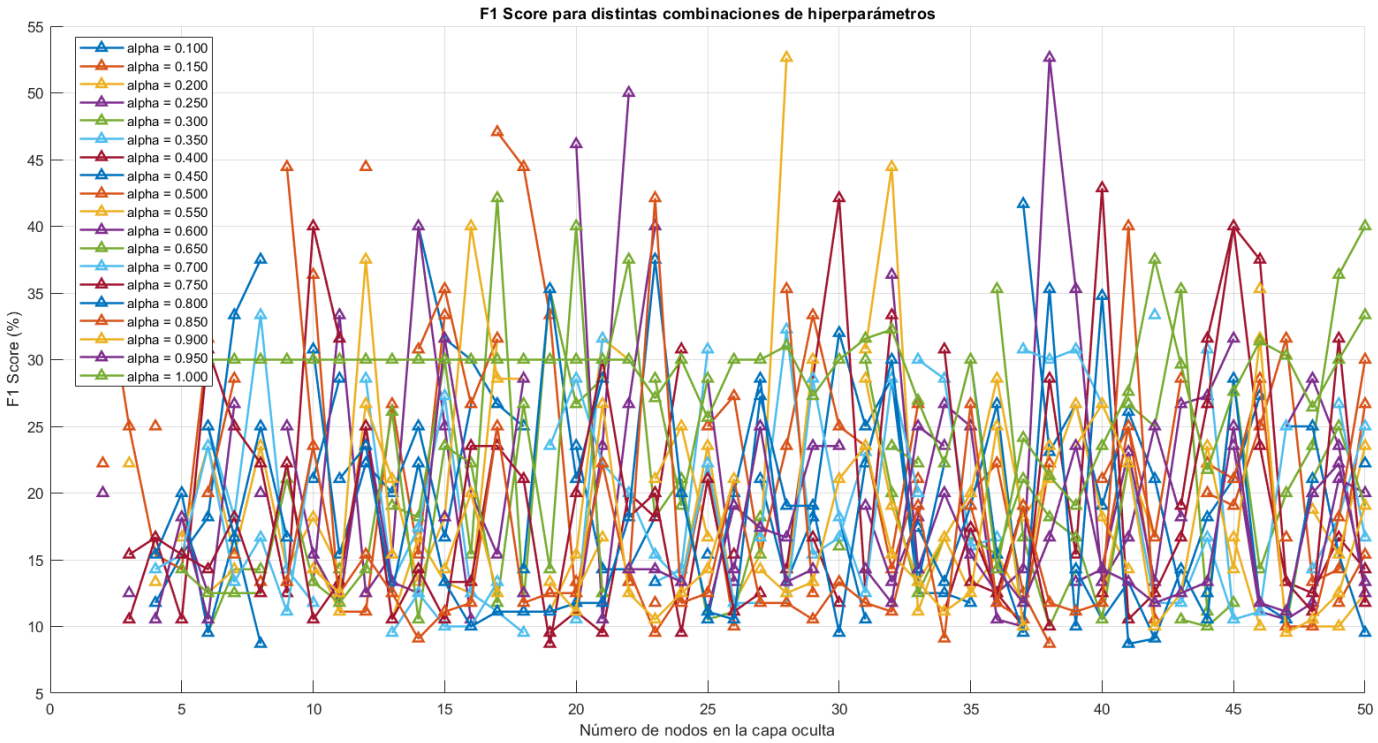

Figura 41. Selección del número de neuronas en la capa oculta y el parámetro de regularización (alpha) que maximizan el *F1 score* en la red neuronal MLP.

5.6. Validación de los modelos

Una vez determinados los hiperparámetros óptimos y haber completado el entrenamiento, se efectuó la validación de los dos modelos generados (RF y red MLP). En el presente apartado, se exponen los resultados de estas validaciones, incluyéndose tanto la validación interna basada en la base de datos retrospectiva, como la validación temporal independiente a partir de la cohorte prospectiva.

5.6.1. *Random Forest*

Validación interna

La Figura 42 muestra la curva ROC del modelo predictivo confeccionado basado en RF y evaluado sobre el conjunto *test* de los datos retrospectivos:

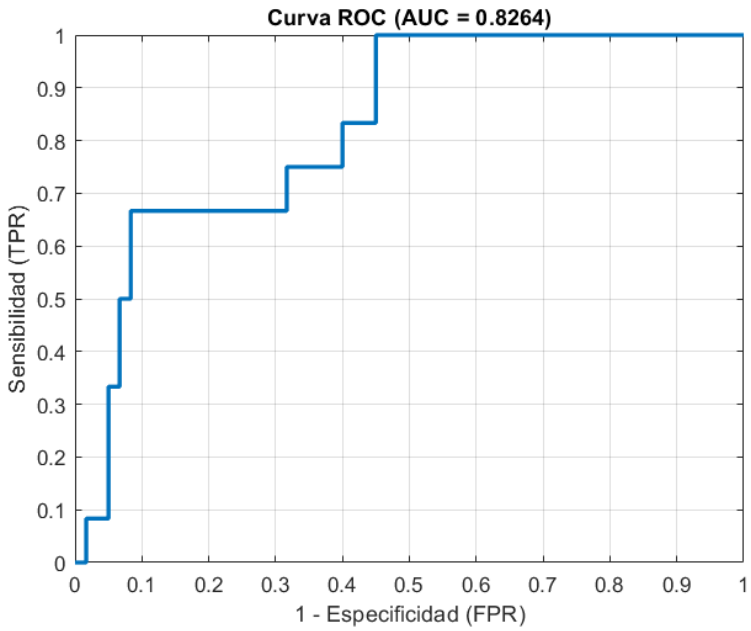

Figura 42. Curva ROC del modelo predictivo basado en *Random Forest* sobre el conjunto *test* de los datos retrospectivos.

A continuación, en la Figura 43 se expone la matriz de confusión resultante al evaluar el rendimiento del modelo en el conjunto *test*:

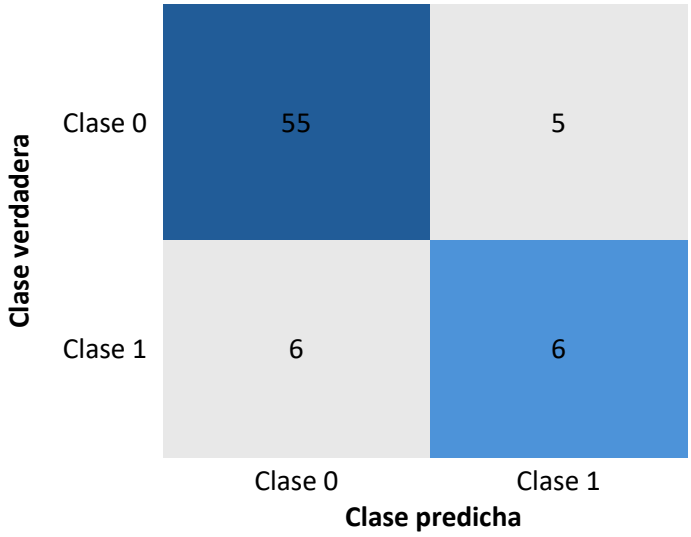

Figura 43. Matriz de confusión para la evaluación interna de la eficacia del modelo basado en *Random Forest* sobre el conjunto *test* de la base de datos retrospectiva.

Se puede apreciar que, de los 60 pacientes correspondientes a la clase negativa, 55 fueron predichos correctamente como no reingreso y 5 fueron clasificados erróneamente como reingreso. Respecto a la clase positiva conformada por 12 pacientes, 6 fueron clasificados erróneamente y otros 6 correctamente.

En la Tabla 32 se resumen las métricas de rendimiento calculadas. Dado el marcado desequilibrio entre clases presente en el conjunto de datos, se observa un notable desbalance entre la sensibilidad (Se) y especificidad (Sp).

Tabla 32. Métricas de rendimiento para la validación interna del modelo predictivo basado en *Random Forest* sobre la base de datos retrospectiva.

Se	Sp	Acc	*F1 score*	LR+	LR-	NPV	PPV	AUC
50.0%	91.7%	84.7%	52.2%	6.00	0.55	90.2%	54.6%	0.826

Se: sensibilidad; Sp: especificidad; Acc: precisión; LR+: razón de verosimilitud positiva; LR-: razón de verosimilitud negativa; NPV: valor predictivo negativo; PPV: valor predictivo positivo; AUC: área bajo la curva ROC.

Validación temporal prospectiva

En la Figura 44 se muestra la matriz de confusión obtenida en esta validación:

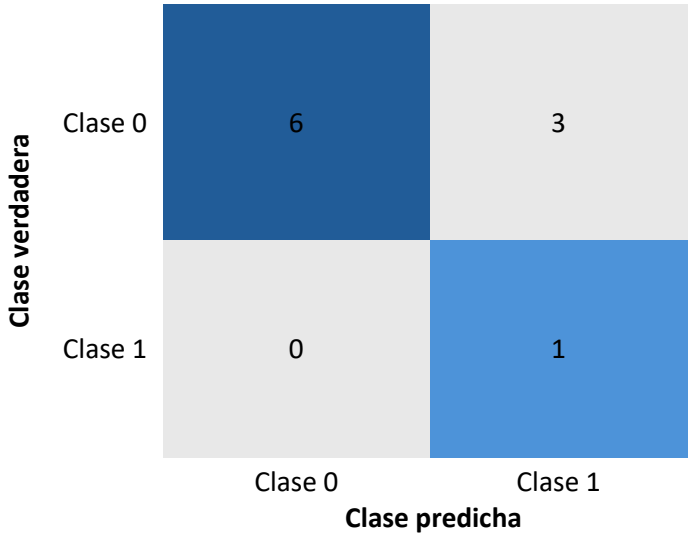

Figura 44. Matriz de confusión para la evaluación de la eficacia del modelo basado en *Random Forest* sobre la base de datos prospectiva.

En ella se observa que, de los 9 sujetos de clase 0 (no reingreso), 6 se clasificaron correctamente y 3 fueron identificados erróneamente como reingresos. Cabe destacar que, aunque esta base de datos dispusiese únicamente de un paciente de clase positiva, el modelo fue aun así capaz de detectarlo correctamente. Si bien puede parecer a simple vista un número muy bajo de casos positivos, encaja con la proporción habitual en el ámbito clínico según la literatura, donde se estima que en torno al 11% de sujetos reingresan.

Por otra parte, la Tabla 33 refleja las métricas de rendimiento obtenidas:

Tabla 33. Métricas de rendimiento para la evaluación del modelo predictivo basado en *Random Forest* sobre la base de datos prospectiva.

Se	Sp	Acc	*F1 score*	LR+	LR-	NPV	PPV
100%	66.7%	70.0%	40.0%	3.00	0.0	100%	25.0%

Se: sensibilidad; Sp: especificidad; Acc: precisión; LR+: razón de verosimilitud positiva; LR-: razón de verosimilitud negativa; NPV: valor predictivo negativo; PPV: valor predictivo positivo.

5.6.2. Red neuronal MLP

Validación interna

La Figura 45 muestra la curva ROC del modelo predictivo confeccionado basado en MLP y evaluado sobre el conjunto *test* de los datos retrospectivos. Asimismo, la Figura 46 constituye la matriz de confusión resultante para la aplicación del modelo sobre esta partición de *test*.

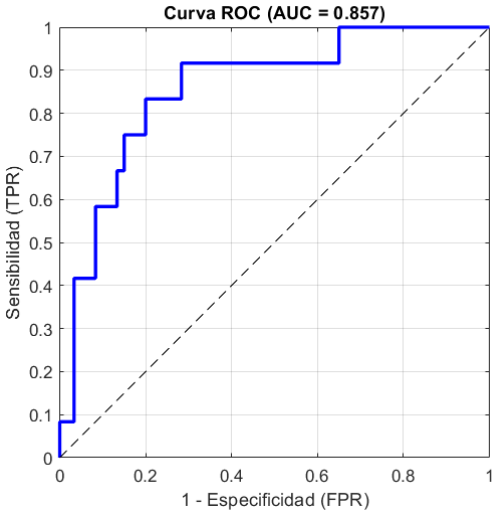

Figura 45. Curva ROC del modelo predictivo basado en una red neuronal MLP sobre el conjunto *test* de los datos retrospectivos.

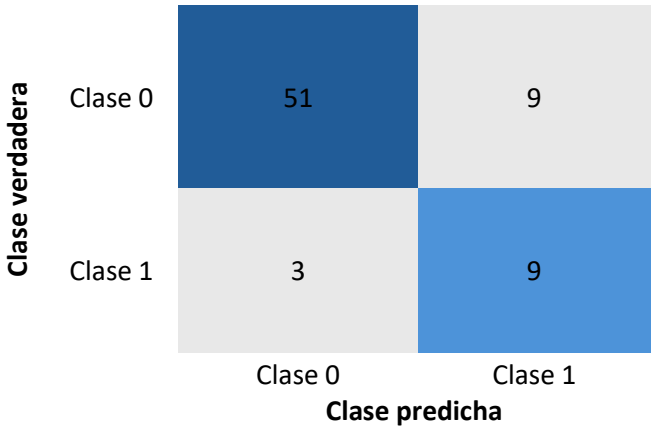

Figura 46. Matriz de confusión para la validación interna de la eficacia del modelo basado en MLP sobre el conjunto *test* de la base de datos retrospectiva.

En esta última puede verse que, del total de 60 pacientes de clase negativa, 51 fueron clasificados correctamente y 9 fueron asignados erróneamente como reingresos. En cuanto a la clase positiva compuesta por 12 pacientes, 9 fueron categorizados adecuadamente como reingresos, mientras que 3 se atribuyeron a la clase negativa sin serlo.

Por otro lado, la Tabla 34 recoge las métricas de rendimiento obtenidas en este caso:

Tabla 34. Métricas de rendimiento para la evaluación interna del modelo predictivo basado en MLP sobre el conjunto *test* de la base de datos retrospectiva.

Se	Sp	Acc	*F1 score*	LR+	LR-	NPV	PPV	AUC
75.0%	85.0%	83.3%	60.0%	5.0	0.3	94.4%	50.0%	0.857

Se: sensibilidad; Sp: especificidad; Acc: precisión; LR+: razón de verosimilitud positiva; LR-: razón de verosimilitud negativa; NPV: valor predictivo negativo; PPV: valor predictivo positivo; AUC: área bajo la curva ROC.

Validación temporal prospectiva

La Figura 47 representa la matriz de confusión resultante:

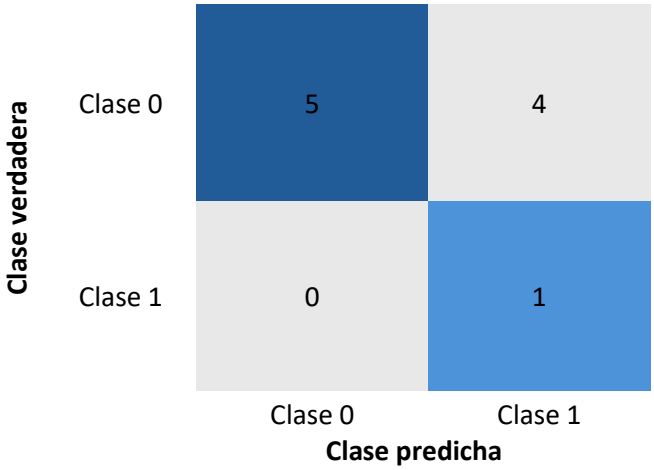

Figura 47. Matriz de confusión para la evaluación de la eficacia del modelo basado en MLP sobre la base de datos prospectiva.

De los 9 pacientes de la clase negativa, 5 fueron clasificados correctamente, mientras que 4 fueron considerados erróneamente como reingresos. Por otra parte, al igual que en la aplicación del conjunto prospectivo en *Random Forest*, el único sujeto de la clase positiva fue identificado adecuadamente.

Finalmente, en la Tabla 35 se exponen las métricas de rendimiento calculadas:

Tabla 35. Métricas de rendimiento para la evaluación temporal del modelo predictivo basado en la red neuronal MLP sobre la base de datos prospectiva.

Se	Sp	Acc	*F1 score*	LR+	LR-	NPV	PPV
100%	55.6%	60.0%	33.3%	2.25	0.0	100%	20.0%

Se: sensibilidad; Sp: especificidad; Acc: precisión; NPV: valor predictivo negativo; PPV: valor predictivo positivo.

CAPÍTULO 6. DISCUSIÓN

Tras la presentación de los resultados obtenidos en la sección anterior, el presente capítulo tiene como objetivo analizar e interpretar los hallazgos principales. Con ello, no solo se pretende alcanzar una mayor compresión de los resultados, sino también proporcionar un enfoque crítico y reconocer posibles limitaciones en el estudio. En primer lugar, se profundizará en la extracción de características y su análisis descriptivo. Tras ello, se abordará la selección de variables, tratando de relacionar la elección automática de las mismas con aspectos clave de la enfermedad que contribuyen a que éstas tengan gran peso en el modelo final. Posteriormente, se abarcará el desempeño predictivo de los modelos automáticos diseñados, estableciendo también una comparativa entre los mismos. Por último, se reflexionará acerca de las posibles limitaciones inherentes al estudio, entendiendo éstas como aspectos que podrían haber condicionado algunos de los resultados expuestos.

6.1. Extracción de características y análisis descriptivo

En el presente apartado, se discutirán las diferencias observadas entre las clases bajo estudio, tanto de la cohorte retrospectiva como prospectiva, comenzando por la primera de estas.

Respecto a las variables sociodemográficas y antropométricas, se observaron diferencias estadísticamente significativas en la variable “Hogar (Residencia)”, con una mayor proporción de pacientes en residencias pertenecientes al grupo de reingreso frente al grupo de no reingreso. Este resultado no es esperable, puesto que, en estos centros, la atención médica continua y el control diario brindado por cuidadores deberían contribuir a disminuir la probabilidad de reingreso hospitalario. Este hallazgo inesperado podría explicarse por posibles comorbilidades severas en estos pacientes. Este resultado guarda relación con las diferencias apreciadas en la categoría “No” de “Cuidador”. Una menor frecuencia de reingreso en sujetos que no disponen de cuidador es un evento inesperado, pero podría justificarse por un posible grado de autonomía elevado en pacientes cuya EPOC está controlada o no se encuentra en fases tardías.

En cuanto a los hábitos y datos clínicos, como se indicó con anterioridad, no se mostraron diferencias significativas. No obstante, cabría esperar diferencia en el hábito e índice tabáquico, ya que este es uno de los principales factores de riesgo de la enfermedad y está estrechamente relacionado con su progresión (como se explicó en el capítulo introductorio).

En referencia a las comorbilidades, el predominio de insuficiencia respiratoria en pacientes de clase positiva frente a los de clase negativa resulta razonable, dado que esta patología es una de las causas más frecuentes de fallecimiento en individuos con EPOC [80]Por tanto, cabe esperar que quienes la padecen reingresen más a menudo. Asimismo, las taquiarritmias son más

frecuentes con la EPOC y sus agudizaciones, lo que explicaría una mayor proporción de reingreso para los que la sufren [81].

Por otra parte, la espirometría previa al ingreso muestra valores significativamente más bajos en "FVC (% teórico)", FEV1 y "FEV1 (% teórico) en la clase positiva frente a la negativa. Este hallazgo es coherente con la fisiopatología de la enfermedad, puesto que, como se indicó en el capítulo introductorio, el curso de la EPOC se caracteriza por una disminución del área funcional de las vías aéreas y la obstrucción de estas. Por tanto, los niveles inferiores de FVC y FEV1 reflejan una mayor severidad de la enfermedad y, por ende, mayor probabilidad de reingreso en pacientes con este perfil clínico.

Por otro lado, los pacientes del grupo de reingreso presentan una mayor proporción de riesgo alto, fenotipo agudizador con enfisema, estadio GOLD D y más ingresos previos por agudización, además de una proporción inferior en el estadio GOLD B frente a la clase negativa. Estos hallazgos son esperables, dado que pacientes con fenotipo agudizador y estadios con mayor carga sintomática pueden relacionarse con una mayor probabilidad de reingreso. Asimismo, como se comentó previamente en este trabajo, los antecedentes de ingresos por exacerbaciones aumentan la probabilidad de ingresar de nuevo.

Respecto a la terapia basal, la mayor aplicación de oxigenoterapia continua domiciliaria y de medicación de rescate con SABA en el grupo de clase positiva (reingreso) respecto a la clase negativa, podría ser un indicativo de mayor severidad de la EPOC. Por tanto, estos eventos son esperables, dado que una mayor gravedad de la enfermedad puede asociarse a exacerbaciones más frecuentes y, en consecuencia, mayor probabilidad de reingreso.

En relación a los tests y cuestionarios realizados, las diferencias significativas vistas en el test de Barthel resultan esperables, ya que una mayor puntuación en el grupo de no reingreso indica mayor autonomía. Asimismo, las puntuaciones más elevadas en el test de CAT para los reingresos y la mayor proporción de pacientes de clase positiva que presentan problemas de movilidad, cuidado personal y relacionados con actividades cotidianas según EuroQol-5D, son eventos coherentes con el deterioro asociado a la enfermedad.

La duración de la estancia hospitalaria en el ingreso presenta también diferencias significativas entre los dos grupos. El hecho de que la duración resulte mayor en los sujetos de clase positiva es un hallazgo que podría dar lugar a dos interpretaciones opuestas. Por un lado, podría relacionarse una estancia más prolongada con una clínica más severa del paciente, lo que justificaría un posterior reingreso. Sin embargo, otro punto de vista es que, una prolongación del número de días hospitalizado se traduce en un mayor periodo en un entorno controlado en el que se brindan cuidados y la atención médica necesaria al sujeto. Esto debería evitar un potencial reingreso *a posteriori*, por lo que el evento podría ser no esperable.

Por otra parte, los pacientes que reingresan presentan diferencias significativas en la PCO_2 (presión parcial de CO_2 en sangre) y bicarbonato al ingreso respecto a los que no reingresan. Los

valores de los sujetos de clase 1 son superiores en los dos casos. Atendiendo a los resultados normales de una gasometría (PCO_2: 38-42 mmHg y HCO_3: 22 -28 mol/L) [82], la mediana de los niveles de PCO_2 en ambos grupos es superior al rango de normalidad, pero mucho mayor en los que finalmente reingresan (43.3 frente a 47 mmHg), manifestando hipercapnia y una mayor severidad de la enfermedad. En cuanto a los niveles de bicarbonato, únicamente la clase positiva presenta valores más elevados de la normalidad, por lo que también implica mayor gravedad de su EPOC.

Respecto a la terapia al ingreso, un mayor número de pacientes de clase positiva muestra un mayor empleo de mucolíticos y ventilación no invasiva frente a la clase negativa. Estos hallazgos pueden tener, al igual que ocurría en la duración de la estancia hospitalaria, una interpretación dual. Por un lado, una aplicación superior de estas estrategias terapéuticas podría traducirse en una mayor severidad de la enfermedad, hecho que explicaría un reingreso posterior. No obstante, algunos estudios han demostrado una reducción del número de exacerbaciones gracias a la administración de mucolíticos [83], mientras que la aplicación de ventilación no invasiva (VNI) contribuye a mejorar el pH y la PCO_2 [84]. Esto podría disminuir las rehospitalizaciones, por lo que el resultado obtenido es parcialmente inesperado.

En cuanto a la terapia al alta, el uso de teofilinas muestra una proporción superior en los reingresos. Este hecho no es esperable, ya que está demostrado que la pauta de teofilinas mejora la función respiratoria y la oxigenación en sujetos con EPOC [85].

Por último, los pacientes que reingresan asisten más a urgencias tras el alta que la clase negativa, lo que podría reflejar un estado de salud inestable.

Por otro lado, en la base de datos prospectiva, el paciente que reingresó presenta un mayor número de ingresos por exacerbación en el año previo. Este hallazgo resulta coherente con lo explicado en el capítulo introductorio y en este apartado, ya que un mayor número de ingresos por agudización se asocia a un mayor riesgo de reingreso y, por tanto, peor pronóstico. Asimismo, la mayor prevalencia de incumplimiento errático según el test TAI en la clase positiva constituye un evento esperable, puesto que alteraciones en la pauta de inhaladores podrían traducirse en un control inadecuado de la enfermedad y, por consiguiente, mayor probabilidad de reingreso. Esto destaca la importancia de la adherencia al tratamiento en la evolución de la patología.

La administración de mucolíticos al ingreso y en estado basal, el empleo de oxigenoterapia basal domiciliaria y un grado de obstrucción al flujo aéreo grave en el sujeto que reingresó, reflejan una mayor severidad de la EPOC. Por tanto, la presencia de estos factores podría ser un indicativo de mayor riesgo de rehospitalización. Sin embargo, como se ha expuesto anteriormente, estudios previos han demostrado una reducción de las agudizaciones mediante la administración de mucolíticos, por lo que su presencia en el reingreso no es completamente esperable. Aun así, es importante tener en cuenta que la clase positiva consta únicamente de 1 paciente, por lo que las interpretaciones extraídas podrían estar sesgadas y no son generalizables.

Finalmente, la ausencia de casos con movimientos torácicos paradójicos, anemia, neumonía, ventilación no invasiva al ingreso y microorganismos resistentes como causa de ingreso podría deberse al reducido tamaño muestral del que se dispone, lo que condiciona su detección.

6.2. Selección de variables

El algoritmo *ReliefF* seleccionó 24 variables predictivas, observándose algunas coincidencias en distintos conjuntos de características al variar *K*. Esto evidencia que las variables finalmente implementadas en los modelos son realmente aspectos determinantes en el desarrollo y curso de la enfermedad, con independencia de los parámetros aplicados.

La pauta de mucolíticos fue la característica a la que *ReliefF* atribuyó una mayor importancia en el estudio. Como se mencionó en el anterior apartado, su administración reduce el número de exacerbaciones. Por tanto, es un predictor clave de reingresos, esperando que su aplicación mitigue considerablemente el riesgo de rehospitalización. Su uso tanto al alta (segunda variable más relevante) como terapia basal destaca su relevancia independientemente del contexto temporal.

La presencia de microorganismos resistentes como motivo de ingreso constituyó la tercera característica con mayor peso. Esta está vinculada a la causa bacteriana (también seleccionada), puesto que ambas resaltan la importancia de las infecciones y los problemas derivados del uso de antibióticos. Estos últimos provocan resistencia en los microorganismos patológicos. Además, como se ha expuesto al comienzo de este trabajo, las infecciones son uno de los procesos fisiopatológicos más característicos de la enfermedad. Por ello, resulta lógico que ambas variables fueran seleccionadas en el estudio.

El test de TAI adopta un papel fundamental en la predicción de reingresos, con el incumplimiento inconsciente como la cuarta variable con más puntuación y el incumplimiento errático también incluido. El valor predictivo de ambos reside en que éstos constatan la adherencia a los inhaladores, por lo que un incumplimiento de la pauta de los mismos conlleva a un descontrol de las manifestaciones de la enfermedad y, por consiguiente, un mayor riesgo de reingresar. Otro test seleccionado por *ReliefF* es el EuroQoL-5D, particularmente la sección que evalúa las dificultades presentadas en el desarrollo de las actividades relacionadas con el cuidado personal. La presencia de limitaciones en el cuidado personal es indicativa de un menor grado de autonomía, contribuyendo a potenciar el evento de reingreso.

Por otra parte, la oxigenoterapia domiciliaria basal es la quinta variable seleccionada. Como se ha explicado en el apartado anterior, así como en la introducción de este estudio, los pacientes con EPOC experimentan una obstrucción al flujo aéreo. En consecuencia, se desencadena un déficit de oxígeno que debe ser solventado mediante el suministro de una fuente externa. Esto justificaría su selección como variable predictora. Además, el hecho de necesitar ser pautada

para algún paciente revela un estadio severo de la enfermedad. Esto refuerza la inclusión de esta característica en los modelos.

También son relevantes para la predicción otras terapias al alta, como las teofilinas (sexta característica seleccionada), cuyo fin es mejorar la oxigenación, como se mencionó previamente. Su uso como terapia basal y al ingreso constituye también otras dos variables seleccionadas y refuerza la importancia de su administración.

El uso de musculatura accesoria (séptima variable seleccionada) está estrechamente ligado con la limitación al flujo aéreo recién comentada. Esta intervención muscular podría explicarse por el trabajo respiratorio adicional efectuado por el paciente como estrategia compensatoria a la falta de oxigenación. Por tanto, el empleo de la misma podría ser un claro predictor de exacerbaciones graves. Esta misma explicación puede extrapolarse a la detección de movimientos torácicos paradójicos, siendo otro reflejo de distrés respiratorio al ingreso que debe ser tratado adecuadamente porque podrían ser claros determinantes de una nueva agudización que motive el reingreso.

Asimismo, el estadio GOLD es otra de las variables predictoras seleccionadas. Esto resulta esperable, ya que aporta información tanto de la carga sintomatológica, como del número de exacerbaciones. Ambos aspectos están correlacionados con un riesgo mayor de reingreso.

La administración de broncodilatadores (LABA y SABA) al ingreso, así como al alta (SABA) fueron también incluidas en el conjunto de entrada de los modelos predictivos. Su uso supone una intervención para disminuir la frecuencia de las exacerbaciones, como se ha expresado en la introducción. Por ello, su inclusión como predictores resulta razonable. Además, como terapia al ingreso, destaca también la aplicación de ventilación no invasiva que mejora los parámetros gasométricos según se ha explicado en el anterior apartado.

La anemia elegida como variable predictora constituye una comorbilidad importante, dado que puede afectar negativamente a diferentes factores de la enfermedad, como la disnea, la tolerancia a la actividad física y la calidad de vida [86]. Por ende, su papel como característica predictora es destacable. Otra comorbilidad importante es la depresión, ya que la disnea provocada por la enfermedad se traduce en un estrés crónico para el paciente que limita sus actividades cotidianas y desencadena en un estado mental alterado. Asimismo, la depresión constituye un predictor de mortalidad para sujetos con EPOC severa [87]. La integración de esta variable en los modelos predictivos ensalza la importancia de los trastornos psicológicos en el curso de la enfermedad, aspecto muchas veces ignorado, pero con un potencial impacto.

Los edemas periféricos al ingreso son uno de los signos de agudización grave [88], además de suponer una clara sospecha de insuficiencia respiratoria en el paciente [89]. Por ello, su inclusión como variable predictora resulta previsible.

Además, la selección de complicaciones por arritmias por parte de *ReliefF* es esperable, ya que, como se indicó en el anterior apartado, la incidencia de las mismas aumenta en las agudizaciones. Por consiguiente, su existencia podría ser un determinante de que se producirán reingresos con mayor probabilidad.

Por último, pero no menos relevante, el número de exacerbaciones por agudización el año previo es una variable que cabía esperar que fuese incluida. Dado que los ingresos guardan relación con la mortalidad, como se explicó en la introducción de este trabajo, un mayor número de exacerbaciones es sinónimo de mayor gravedad de la enfermedad y, en consecuencia, mayor riesgo de volver a ingresar.

6.3. Evaluación de la predicción de los modelos y comparativa entre ambos

En esta sección, se analizan los resultados obtenidos por el modelo predictivo basado en *Random Forest,* así como por la red neuronal MLP. Con este fin, la estructura de este apartado es la siguiente: en primer lugar, se evaluará el rendimiento predictivo de RF tanto sobre la base de datos retrospectiva como prospectiva; tras esto, se procederá a realizar un análisis análogo para MLP y, finalmente, se establecerá una comparativa entre ambos modelos predictivos tratando de concluir cuál de ellos presenta una mayor capacidad predictiva.

6.3.1. *Random Forest*

Los resultados de este modelo predictivo en ambos conjuntos de datos muestran un desempeño sólido. Sin embargo, es importante destacar ciertos aspectos ocasionados por la naturaleza desbalanceada de ambos grupos.

En el conjunto de *test* de la **base de datos retrospectiva** (validación interna), la especificidad alcanzada fue alta (91.7%), mientras que la sensibilidad resultó ser moderada (50.0%). Esto puede traducirse en una capacidad elevada para clasificar correctamente los pacientes del grupo no reingreso, pero algo comprometida para detectar los de clase positiva. Por su parte, la *F1 score* (52.2%) refleja una capacidad moderada para distinguir entre grupos, dado que la sensibilidad se ve algo disminuida. Este hecho podría deberse al gran desequilibrio entre clases del *dataset* de entrenamiento *train2*, ya que se disponen de 21 reingresos y 99 no reingresos, lo que condiciona el aprendizaje de patrones de pacientes de clase positiva. En la matriz de confusión (Figura 43) puede apreciarse cierta dificultad para identificar casos de reingresos (de 12 casos positivos, se detectan 6). No obstante, el valor de AUC obtenido es de 0.826, lo que demuestra que, a pesar del desbalance, el modelo tiene una capacidad predictiva destacable.

Respecto a la **base de datos prospectiva** (validación temporal), un aspecto positivo a destacar es que el modelo fue capaz de identificar el único paciente de clase positiva, motivo por el que su sensibilidad fue del 100%. Sin embargo, su especificidad resultó ser del 66.7%, apreciándose 3 falsos positivos en su matriz de confusión (Figura 44). Esto podría ser un indicativo de sobreestimación de los casos positivos, conclusión reforzada por el valor reducido de PPV (25.0%). Dado que la *F1 score* es una métrica dependiente de la sensibilidad y el PPV, esta se ve algo reducida (40.0%).

En **ambas bases de datos**, el modelo basado en *Random Forest* logra un potencial predictivo satisfactorio. La elevada especificidad (91.7%) y el valor de AUC próximo a la unidad (0.826) en la base retrospectiva, avalan su notable desempeño y destaca su detección adecuada de pacientes de clase negativa (no reingresos). En contraste, en la base de datos prospectiva, a pesar de su reducida población de estudio (10 pacientes), se alcanza la máxima sensibilidad posible (100%), subrayando su superioridad para identificar casos positivos respecto a la retrospectiva. Sin embargo, las dimensiones tan reducidas de su muestra limitan métricas como la *F1 score* (40.0%) y la precisión (70%), siendo esta última inferior a la obtenida en los datos retrospectivos (84.7%).

Un aspecto notable es precisamente el comportamiento opuesto entre las métricas de sensibilidad y especificidad en las dos bases de datos. La cohorte retrospectiva posee una sensibilidad menor que su especificidad, mientras que la prospectiva consta de una sensibilidad superior a su especificidad. Esto podría atribuirse a las diferencias en las características de las cohortes, puesto que la retrospectiva se recopiló antes de la pandemia de COVID-19 y la prospectiva posteriormente. Por tanto, todos los cambios organizativos y la implementación de nuevos protocolos que tuvieron lugar entre las adquisiciones de datos pudieron suponer un factor introductor de variabilidad entre ellas.

En términos generales, la base de datos retrospectiva presenta un mayor rendimiento, además de constar de un mayor número de pacientes que hace sus resultados más generalizables. A pesar de esto, dado que la base de datos prospectiva es aplicada en este trabajo como validación temporal independiente del modelo generado con la retrospectiva, esta confirma un funcionamiento satisfactorio de *Random Forest* en nuevos conjuntos de datos, observando su habilidad para identificar pacientes que reingresan incluso cuando los tamaños muestrales son tan disminuidos.

6.3.2. Red neuronal perceptrón multicapa

En el conjunto de *test* de la **base de datos retrospectiva** (validación interna), el modelo refleja un rendimiento satisfactorio, evidenciado por su sensibilidad (75.0%), especificidad (85.0%) y precisión (83.3%). Estos valores muestran su capacidad para clasificar adecuadamente tanto a los pacientes de clase positiva, como a los de clase negativa. En su matriz de confusión (Figura

46), se aprecian valores reducidos de falsos positivos (9 sobre 60) y negativos (3 de 12) en comparación con la clase verdadera total, lo que constituye un indicio más de su funcionamiento satisfactorio en la tarea predictiva.

Asimismo, el *F1 score* fue del 60.0%, lo que implica un equilibrio moderado entre la sensibilidad del modelo y su valor predictivo positivo o PPV (50.0%). Este último sugiere que la mitad de las predicciones positivas del modelo eran realmente reingresos, por lo que podrían aplicarse estrategias para optimizar este aspecto. No obstante, este hecho puede atribuirse al desbalance de clases, siendo mayoritaria la clase negativa y dificultando el aprendizaje del grupo positivo.

Por otra parte, su valor predictivo negativo (NPV) del 94.4% revela que se clasificaron adecuadamente al 94.4% de sujetos del grupo de no reingreso. Además, el AUC fue de 0.857, lo que se traduce en una buena capacidad predictiva del modelo.

En cuanto a la **base de datos prospectiva** (validación temporal), se alcanzó una sensibilidad del 100%, puesto que la red pudo identificar el único paciente perteneciente a la clase positiva. No obstante, su especificidad (55.6%) revela ciertas dificultades para predecir los casos negativos. Esto es avalado por la matriz de confusión resultante (Figura 47), en la que puede observarse un total de 4 falsos positivos sobre 9 pacientes de clase negativa. Su PPV del 20.0% indica una capacidad baja de predecir correctamente los reingresos, a pesar de sí haberlo hecho para el único paciente del que se dispone. Esto ocasiona cierta desconfianza en sus predicciones positivas, aunque podría estar motivado por el disminuido número de reingresos del conjunto. Un aspecto destacado es que todas las predicciones negativas (no reingreso) fueron correctas, hecho que justifica su NPV del 100%.

Dado que en este estudio la base de datos prospectiva se integra como validación temporal independiente, es importante destacar que, a pesar de sus limitaciones por la baja disponibilidad de datos, métricas como el NPV (100%) y la sensibilidad (100%) muestran valores muy satisfactorios. Estos reflejan una capacidad prometedora de generalización en nuevas bases de datos, pero con matices de mejora como implementaciones en conjuntos más grandes que permitan optimizar el rendimiento predictivo del modelo.

Por último, al igual que ocurría en RF, se observa un comportamiento opuesto entre la sensibilidad y especificidad en las bases de datos, reflejando la cohorte retrospectiva (anterior al COVID-19) una menor sensibilidad que especificidad, y la prospectiva (post COVID-19) lo inverso. Como se ha expuesto con anterioridad, esto podría deberse a los cambios organizativos y protocolos entre ambas, lo que aporta cierta variabilidad.

6.3.3. Comparativa entre *Random Forest* y MLP

Pese a que la red neuronal perceptrón multicapa fue diseñada como modelo de referencia frente al principal (*Random Forest*), en la base de datos retrospectiva ha demostrado un rendimiento predictivo global superior (ver Tablas 32 y 34). Su elevada sensibilidad (75.0%) destaca sobre la obtenida en *Random Forest* (50.0%), demostrando una mayor capacidad de detección de casos positivos, un aspecto crítico en el ámbito clínico para adoptar medidas preventivas. Asimismo, en la red neuronal se mitiga en mayor medida el desbalance entre sensibilidad y especificidad observado en RF*,* derivado del gran desequilibrio entre clases. Esto indica que este modelo maneja de manera más optima conjuntos de datos con una clase mayoritaria muy por encima de la minoritaria como es este caso.

Por otra parte, el *F1 score* en esta cohorte retrospectiva es superior en el MLP (60.0% frente a 52.2%), por lo que en este modelo existe un mayor equilibrio entre sensibilidad y valor predictivo positivo (PPV). No obstante, el PPV es ligeramente superior en *Random Forest* (54.6% frente a 50.0%), lo que se traduce en una fiabilidad levemente superior en sus predicciones positivas. Por tanto, al tratarse de una discrepancia tan pequeña, este hallazgo no es un indicio claro de mejor capacidad predictiva de RF.

Este último, en contraste con la red neuronal, presenta una especificidad superior (91.7% frente a 85.0%), lo que significa que posee una capacidad elevada para identificar los pacientes del grupo de no reingreso. Este aspecto permite evitar falsos positivos, lo que puede extrapolarse a una menor inversión de recursos clínicos innecesarios.

La precisión (*accuracy*) no muestra prácticamente discrepancias entre ambos modelos, obteniéndose resultados muy similares (84.7% en *Random Forest* y 83.3% en el MLP). Esto destaca la eficacia de los dos diseños ya que sugiere una elevada proporción de predicciones correctas sobre el total de predicciones efectuadas. Sin embargo, el AUC en la red neuronal (0.857) destaca sobre el calculado en *Random Forest* (0.826)*,* lo que sugiere que la capacidad predictiva del MLP resulta superior. Esto no excluye a RF de ser un modelo eficiente, ya que valores superiores a 0.8 reflejan una buena capacidad discriminativa entre clases.

En resumen, el rendimiento del MLP en la cohorte retrospectiva resulta superior, lo que se evidencia en métricas como la sensibilidad, *F1 score* y AUC. Por consiguiente, resulta más efectivo en la detección de pacientes que reingresan, un aspecto clave en este trabajo dado que la clase minoritaria, con bastante discrepancia, es la positiva. No obstante, *Random Forest* posee una capacidad superior en cuanto a la predicción de no reingresos, dada su elevada especificidad, por lo que evita falsos positivos.

El motivo por el que MLP ofrece un mejor funcionamiento que RF podría residir en que este último dispone de un mayor número de hiperparámetros a optimizar. A mayor número de hiperparámetros, más instancias o pacientes se requieren para lograr una optimización global adecuada, lo que se traduce en un modelo con un desempeño superior. Dado que la red MLP

consta únicamente de tres hiperparámetros a ajustar (número de neuronas en la capa oculta, parámetro de regularización y umbral de decisión), la optimización resultó más efectiva con el mismo número de pacientes.

Respecto a la base de datos prospectiva, pese a que ambos modelos alcanzaron una sensibilidad del 100%, RF posee una especificidad superior (66.7% frente a 55.6% en la red MLP). Esto es indicativo de una mayor capacidad de identificar a los pacientes de clase negativa, mitigando los falsos positivos. Este hallazgo se ve reforzado por su PPV mayor (25.0% frente a 20.0% en MLP).

En cuanto a la precisión, RF también demuestra su superioridad con un valor del 70.0% en comparación con el 60.0% obtenido en la red neuronal. Por tanto, RF ofrece una mayor proporción de predicciones correctas sobre el total de predicciones realizadas. Además, la métrica *F1 score* resulta más adecuada en el caso de RF (40.0% frente a 33.3% en MLP), por lo que este modelo demuestra un mayor equilibrio entre la sensibilidad y el valor predictivo positivo (PPV).

Otro punto a destacar es que ambos modelos logran un NPV del 100%, es decir, todos los casos predichos como negativos resultaron ser finalmente no reingresos. Por tanto, tanto RF como la red MLP reflejan una gran habilidad para descartar falsos negativos.

En definitiva, *Random Forest* posee una capacidad de generalización superior al implementarse sobre conjuntos de datos nuevos, aspecto que puede observarse en métricas como la especificidad, *F1 score*, PPV y precisión. Por consiguiente, el modelo basado en RF demuestra una mayor robustez y, en consecuencia, mayor fiabilidad de sus predicciones en entornos clínicos.

6.4. Comparación con otros estudios

La predicción de reingresos hospitalarios por exacerbación de EPOC mediante técnicas de *Machine Learning* ha sido abordada ampliamente en la literatura. Existe una gran diversidad de técnicas para construir los modelos, más allá de *Random Forest* y el perceptrón multicapa (MLP). Este gran abanico de oportunidades brinda la posibilidad de adoptar distintas estrategias y configuraciones que maximicen el rendimiento según el contexto específico.

Se han identificado varias coincidencias en las variables utilizadas respecto a otras investigaciones previas tales como la edad, el sexo, test de Charlson, neumonía como comorbilidad, duración de la estancia hospitalaria, número de exacerbaciones previas, aplicación de broncodilatadores (LAMA, SAMA, LABA y SABA), uso de corticoides, diabetes, hipertensión, presión arterial sistólica y diastólica, frecuencia respiratoria y cardíaca, resultados de analíticas y espirometría, grado de disnea y test de CAT, entre otras [90], [91]. Aunque no todas ellas fueron seleccionadas por *ReliefF*, el hecho de que existan variables comunes con otros estudios refuerza la validez metodológica del presente trabajo.

El estudio de Lin et al. (2024) [90] propone diversos modelos predictivos de readmisión hospitalaria en los 30 días posteriores al alta mediante regresión logística, *Random Forest*, *Extreme Gradient Boosting* (XGBoost) y una red neuronal. Su base de datos incluyó 101011 pacientes retrospectivos y 17565 prospectivos, tamaños muestrales significativamente mayores que los de este TFG. El número de exacerbaciones en el año previo fue una de las variables predictoras clave, al igual que resultó serlo en este trabajo.

Su conjunto retrospectivo proporciona valores inferiores a los obtenidos aquí, independientemente de la técnica de *Machine Learning* propuesta. En concreto, sus métricas de AUC resultaron ser 0.706 para *Random Forest* y 0.707 para el MLP. Respecto a la base de datos prospectiva, *Random Forest* mostró una AUC de 0.714, sensibilidad de 62.0%, especificidad de 69.3%, PPV de 43.4% y NPV de 82.7%. En comparación con las métricas prospectivas de este TFG, las de Lin et al. (2024) [90] muestran una sensibilidad, NPV y AUC inferior, pero mayor especificidad y PPV. La red neuronal de Lin et al. (2024) [90] obtuvo una AUC del 0.707, también inferior a la de este trabajo. Además, el conjunto prospectivo reflejó una AUC del 0.717, sensibilidad de 68.0%, especificidad de 64.1%, PPV de 41.9% y NPV de 84.1%. Esto implica que sus valores fueron inferiores a los obtenidos en este TFG en términos de sensibilidad, AUC y NPV, pero superiores en especificidad y PPV. Sin embargo, la limitada base de datos prospectiva de este TFG reduce la representatividad de las métricas frente al elevado tamaño muestral de Lin et al. (2024) [90]. Asimismo, este artículo integra un enfoque adicional para predecir la mortalidad temprana de estos pacientes.

Por su parte, Chen et al. (2021) [91], analizaron una cohorte de 650 sujetos, reducida a 636 tras la aplicación de criterios de inclusión y exclusión. Los autores propusieron dos modelos predictivos (regresión logística y XGBoost) para predecir el reingreso de pacientes por exacerbación de EPOC. Este artículo propone dos enfoques diferentes: en primer lugar, las técnicas de *Machine Learning* implementadas difieren de las desarrolladas en este TFG, mostrando alternativas funcionales. Por otra parte, en lugar de centrar la predicción en los 30 días posteriores al alta, el artículo abarca un periodo de un año.

El modelo logístico alcanzó unos resultados de 66.7% de sensibilidad; 66.4% de especificidad; precisión de 66.5% y AUC de 0.699. El modelo XGBoost obtuvo una sensibilidad de 63.5%; especificidad de 75.0%; precisión de 71.2% y AUC de 0.722. Comparando estas técnicas con las desarrolladas en este trabajo, ambos modelos superan al *Random Forest* solo en sensibilidad, mostrando inferioridad en el resto de las métricas. Además, el MLP de este TFG supera a ambos modelos en todos los aspectos. Esto refuerza la robustez de los modelos desarrollados en el presente TFG, ya que, a pesar de disponer de una cohorte limitada, ha logrado alcanzar rendimientos superiores a los dados mediante bases de datos más grandes.

Finalmente, López et al. (2023) [92], incluyeron 1905 pacientes con EPOC pertenecientes a una cohorte retrospectiva como entrada a los modelos predictivos. En este caso, la readmisión a los 30 días posteriores al alta constituye una variable predictora, en lugar de ser la variable *target*.

Sin embargo, los modelos propuestos tratan de predecir readmisiones, pero sin especificar un periodo temporal concreto a diferencia de este trabajo. Este estudio introduce técnicas no tratadas en los anteriores ejemplos, como *Naive Bayes* y SVM (*Support Vector Machine*).

Naive Bayes alcanzó una sensibilidad del 13.0%, especificidad del 97.0%, precisión del 79.0% y AUC del 0.830. En comparación con el modelo *Random Forest* de este TFG, *Naive Bay*es ofreció mayor especificidad y su AUC fue levemente superior (0.830 vs. 0.826). No obstante, para el resto de las métricas, *Random Forest* resultó predominante. Por su parte, la red neuronal MLP de este trabajo fue superior en todos los aspectos al de *Naive Bayes*, salvo en la métrica de especificidad.

SVM ofreció una sensibilidad del 62.0%, especificidad del 88.0%, precisión o *accuracy* del 80.0% y AUC del 0.88. Esta técnica resultó superior en términos de sensibilidad y AUC respecto al modelo basado en *Random Forest*, y también en especificidad y AUC en comparación al MLP. Estas superioridades podrían explicarse por la diferencia sustancial en las dimensiones de la población bajo estudio. Además, las diferencias en AUC son pequeñas, especialmente en la red neuronal (0.857 frente a 0.888 de SVM).

Además, los autores también entrenaron sendos modelos predictivos basados en RF y en redes neuronales MLP. El modelo *Random Forest* de López et al. mostró una sensibilidad del 58.0%, especificidad del 90.0%, precisión de 79.0% y AUC del 0.87. Estos resultados son superiores en términos de sensibilidad y AUC al modelo RF desarrollado en este trabajo. No obstante, la discrepancia en la sensibilidad es leve (50.0% en el diseñado y 58.0% en el proporcionado en la publicación). Su red neuronal perceptrón multicapa también resultó superior en las dos métricas citadas, aunque solo ligeramente. Su AUC fue del 0.87, mientras que la obtenida en el presente TFG es de 0.857. Además, su sensibilidad fue del 84.0% frente al modelo RF propuesto en este trabajo, del 75.0%.

Las diferencias observadas con otros estudios similares presente en el estado del arte sugieren que, con una mayor cohorte, los modelos planteados en este TFG potenciarían su capacidad predictiva, puesto que ya resultan competitivos ante investigaciones basadas en conjuntos de datos superiores.

En la Tabla 36, se sintetiza la comparación del presente estudio con las investigaciones recién descritas, permitiendo apreciar de una forma más visual las diferencias entre ellas.

Tabla 36. Comparativa del estudio actual con otras publicaciones científicas similares.

Autor	Año	Tamaño población	Periodo para considerar reingreso	Tipo de variables	Sensibilidad (Se)	Especificidad (Sp)	AUC
Lin et al. [90]	2024	Retrospectiva: 101011 Prospectiva: 17565	30 días	Demográficas Clínicas Condiciones ambientales Analíticas Comorbilidades Antropométricas Terapia al ingreso	Retrospectiva: - RF: s/n - MLP: s/n Prospectiva: - RF: 62.0% - MLP: 68.0%	Retrospectiva: - RF: s/n - MLP: s/n Prospectiva: - RF: 69.3% - MLP: 64.1%	Retrospectiva: - RF: 0.706 - MLP: 0.707 Prospectiva: - RF: 0.714 - MLP: 0.717
Chen et al. [91]	2021	636	365 días	Demográficas Terapia clínica Antecedentes Cuestionarios Analíticas	Regresión logística: 66.7% XGBoost: 63.5%	Regresión logística: 66.4% XGBoost: 75.0%	Regresión logística: 0.699 XGBoost: 0.722
López et al. [92]	2023	1905	Sin especificar	Demográficas Antropométricas Sociales Comorbilidades Terapia al ingreso Analíticas Antecedentes	Naive Bayes: 13.0% SVM: 62.0%	Naive Bayes: 97.0% SVM: 88.0%	Naive Bayes: 0.830 SVM: 0.880
Tamayo Polo. M. (actual TFG)	2025	Retrospectiva: 243 Prospectiva: 10	30 días	Terapia al ingreso Terapia al alta Cuestionarios Terapia Basal Comorbilidades Gravedad enfermedad Causas ingreso	Retrospectiva: - RF: 50.0% - MLP: 75.0% Prospectiva: - RF: 100.0% - MLP: 100.0%	Retrospectiva: - RF: 91.7% - MLP: 85.0% Prospectiva: - RF: 66.7% - MLP: 55.6%	Retrospectiva: - RF: 0.826 - MLP: 0.857 Prospectiva: - RF: s/n - MLP: s/n

s/n: sin número.

6.5. Limitaciones del estudio

Todo estudio está sujeto a limitaciones, por lo que el presente trabajo no está exento de ellas. Esta sección resulta muy relevante, pues permite identificar ciertos puntos débiles que podrían condicionar los resultados finales obtenidos.

Una de las principales limitaciones es su diseño unicéntrico, es decir, haber sido desarrollado en un único centro hospitalario (Hospital Universitario Río Hortega de Valladolid). Esto compromete la generalización del modelo, ya que las variables que conforman la base de datos pueden adquirir valores diferentes en función de distintos contextos geográficos, demográficos o

sociales. No obstante, un único centro de recogida de datos favorece un mayor control en la elaboración de las cohortes.

Otro aspecto destacable son las reducidas dimensiones de las bases de datos, especialmente de la prospectiva conformada tan solo por 10 sujetos, lo que contribuye a una reducción de la robustez de los resultados. Asimismo, las cohortes presentan un desbalance marcado entre la clase positiva (reingreso) y la negativa (no reingreso), potenciando el sesgo hacia la clase mayoritaria (no reingreso en este caso) y dificultando la detección de los casos positivos.

Por otra parte, la variabilidad derivada de los cambios organizativos y de protocolos entre la base retrospectiva (pre COVID-19) y prospectiva (post COVID-19), responsable de diferencias en las características de estas, condiciona la comparabilidad de los resultados.

En cuanto a la recogida de datos prospectivos, el periodo temporal en el que esta tuvo lugar (enero a junio de 2025) coincidió con un descenso en la frecuencia de ingresos por exacerbación de EPOC, dado que estos suelen disminuir a medida que se abandonan las épocas invernales. Esto pudo infrarrepresentar las agudizaciones, limitando el tamaño de la cohorte.

Por otro lado, las predicciones de reingreso se centraron en un periodo de 30 días tras el alta, imposibilitando la extrapolación de los resultados a posibles ingresos tardíos. Asimismo, el estudio se limitó a dos enfoques de aprendizaje computacional (*Random Forest* y MLP), lo que sugiere la posibilidad de aplicar otras estrategias en trabajos futuros.

CAPÍTULO 7. CONCLUSIONES

La Enfermedad Pulmonar Obstructiva Crónica (EPOC) es actualmente la cuarta causa de muerte en el mundo, pudiendo alcanzar el tercer puesto en 2030. Su elevada mortalidad, discapacidad asociada y gran carga económica y asistencial evidencian que se trata de un problema de salud prioritario. Este hecho subraya la imperatividad de adoptar estrategias que contribuyan a su prevención, finalidad buscada en este estudio a través del desarrollo de modelos predictivos de reingresos por exacerbación de EPOC en los 30 días posteriores al alta. De esta forma, resultaría posible implementar medidas preventivas, optimizar la gestión de la enfermedad y mitigar su impacto tanto en la calidad de vida de los pacientes, como en el sistema sanitario.

Para la realización de este estudio, se elaboraron dos bases de datos (una retrospectiva y otra prospectiva) procedentes del Servicio de Neumología del Hospital Universitario Río Hortega de Valladolid. La población retrospectiva constó de 246 sujetos, mientras que la prospectiva de 75. Esta última fue recogida para la validación temporal independiente de los modelos diseñados. Entre estos pacientes, se distinguen dos grupos: clase positiva (reingreso) y negativa (no reingreso). Las variables recogidas fueron de diversa naturaleza, tratando de capturar la enorme heterogeneidad de la enfermedad.

La metodología implementada se basó en cuatro etapas diferentes: (i) *data curation*, (ii) selección automática de las variables predictoras, (iii) diseño y optimización de modelos predictivos de clasificación binaria (reingresos vs. no reingreso); y (iv) validación independiente de los modelos, tanto interna como temporal. La selección de variables predictoras se efectuó mediante el algoritmo *ReliefF*, mientras que la clasificación se basó en dos modelos predictivos: *Random Forest* (RF) y red neuronal artificial perceptrón multicapa (MLP).

Por último, se estableció una comparativa entre el rendimiento de ambos modelos y de estos con los resultantes de investigaciones similares previas en el contexto de la predicción de reingreso por exacerbación de EPOC. De este modo, se evidenció una ligera superioridad de la red neuronal MLP con respecto al modelo RF. En comparación con los resultados de otros estudios, se demostró la superioridad de los modelos desarrollados con respecto a algunos estudios que presentaban cohortes con dimensiones mucho mayores en comparación a las empleadas en este TFG.

7.1. Contribuciones

Las principales contribuciones que derivan de este trabajo son las siguientes:

- Análisis exhaustivo de una base de datos retrospectiva conformada por decenas de variables de caracterización de la enfermedad de EPOC, desde el estado basal del paciente hasta la gravedad de la exacerbación, pasando por el estado al ingreso y al alta. Esto permitió caracterizar estadísticamente las mismas y conocer la influencia de cada una de ellas en el evento de reingreso.
- Selección e identificación de las variables predictoras óptimas a través del algoritmo *ReliefF*, constituyendo un conjunto de características multifactoriales que refleja la heterogeneidad de la enfermedad y la importancia de considerar la misma.
- Desarrollo de nuevos modelos predictivos basados en *Random Forest* y red MLP con AUC superiores a 0.8 en ambos casos, cuyo fin es predecir reingresos por exacerbación de EPOC en los 30 días post-alta y favorecer el ahorro de recursos.
- Validación temporal independiente en una cohorte prospectiva, destacando la importancia de efectuar las validaciones en conjuntos de grandes dimensiones para garantizar una generalización fiable.

7.2. Principales conclusiones del estudio

A continuación, se exponen las principales conclusiones derivadas de este trabajo:

1. El conjunto de variables seleccionadas ha demostrado ser útil para identificar los patrones relacionados con el reingreso por exacerbación de EPOC, validándose la hipótesis clínica. Destacan la pauta de mucolíticos al ingreso y alta, la presencia de microorganismos resistentes, el test TAI de uso correcto de inhaladores y la prescripción de oxigenoterapia domiciliaria basal. Además, la naturaleza diversa de las variables ensalza la importancia de implementar enfoques de análisis de datos multimodales debido a los múltiples factores que influyen sobre el curso de la enfermedad.

2. Los modelos predictivos basados en técnicas de *Machine Learning*, en particular *Random Forest* y red neuronal MLP, demostraron ser efectivos en la predicción de reingresos por exacerbación de EPOC en los 30 días posteriores al alta. Las métricas de rendimiento reflejaron una AUC de 0.826 para *Random Forest* y de 0.857 para la red MLP. El buen desempeño de los modelos en la base de datos retrospectiva confirma la hipótesis técnica establecida.

3. La validación temporal independiente con la base de datos prospectiva confirmó la viabilidad de los modelos en nuevas cohortes. No obstante, dado su reducido tamaño muestral, resultaría esencial aplicar un conjunto de datos más grande para esta validación, consolidando la fiabilidad de las métricas de rendimiento.

4. La comparativa establecida con otros estudios reveló una capacidad predictiva competitiva y, en ciertos casos, superior a estudios que disponen de cohortes de mayor tamaño. Esto sugiere un diseño adecuado y robusto con gran potencial para futuras aplicaciones a pesar del tamaño muestral limitado.

7.3. Líneas futuras de investigación

De acuerdo con las limitaciones expuestas en el anterior capítulo, a continuación, se presentan posibles líneas futuras de investigación que permitan integrar mejoras y contribuir a un mayor alcance del estudio.

Dado que las dimensiones de la cohorte prospectiva son limitadas y el desbalance de clases es muy marcado, el estudio podría beneficiarse de la aplicación de técnicas como SMOTE. Esto posibilitaría la creación de muestras sintéticas de la clase reingreso (minoritaria), aportando una mayor representabilidad de la misma sobre el conjunto de datos y contribuyendo a aumentar la sensibilidad de los modelos.

En relación con este último aspecto sobre el tamaño reducido de los datos prospectivos, puesto que estos se emplean para la validación temporal de los modelos, futuras validaciones independientes podrían incluir cohortes más grandes, permitiendo una validación más fiable.

La incorporación al estudio de un mayor número de centros hospitalarios podría suponer un avance significativo, puesto que evitaría sesgos relacionados con diversos contextos del paciente, permitiría la disponibilidad de bases de datos más amplias y, por consiguiente, potenciaría la generalización de los modelos predictivos desarrollados.

Teniendo en cuenta su satisfactorio desempeño a pesar de las limitaciones en cuanto a disponibilidad de datos, su aplicación podría extenderse a periodos temporales superiores a los 30 días al alta. En este sentido, resultaría valioso explorar y comparar diferentes periodos de seguimiento a corto (semanas), medio (meses) y largo plazo (1 año). Con ello, sería posible identificar el marco temporal óptimo para la implementación de medidas preventivas que permitan optimizar recursos y preservar la calidad de vida de los pacientes.

Por otra parte, podría ser interesante la evaluación del rendimiento de otros enfoques predictivos, como SVM, métodos de *ensemble learning* diferentes a RF tales como AdaBoost, e incluso arquitecturas de *Deep Learning*, en el supuesto de disponer de suficientes datos.

De cara a su implementación en el ámbito clínico, resultaría de gran impacto la integración de técnicas de *Explainable Artificial Intelligence* (XAI). Esto condicionaría positivamente la aceptación de los modelos por parte del personal sanitario dada su escasa familiarización con estas herramientas. Además, permitiría alcanzar la explotación máxima de los modelos, aprovechando las prestaciones que proporciona.

ANEXOS

En la siguiente sección, se adjunta el código desarrollado en las distintas etapas del presente trabajo. Esto incluye el análisis estadístico de las variables, tratamiento e imputación de datos perdidos, selección de características predictoras y los distintos modelos predictivos elaborados junto a su evaluación y validación independiente.

ANEXO 1. Código tratamiento e imputación de datos faltantes.

En primer lugar, se adjunta el código desarrollado para eliminar variables identificativas, relacionadas con el reingreso, fechas o incluso aquellas que fueron recogidas para facilitar la construcción de la base de datos, pero que no tienen ningún tipo de valor predictivo. Se excluye también a aquellos pacientes cuyo alta fuese motivado por su *exitus* y, por tanto, no se pudo realizar el seguimiento de 30 días posteriores al alta como se tenía establecido. Además, se calcula el porcentaje de datos perdidos por fila (paciente) y columna (variable):

```
%-------------------------------------------------------------------------
%-------------------------------------------------------------------------
%          EXPLORACIÓN INICIAL DE LA BASE DE DATOS RETROSPECTIVA
%-------------------------------------------------------------------------
%-------------------------------------------------------------------------

% En primer lugar, se calcula el porcentaje de NaNs para cada fila
(paciente)
% y cada columna (variable) sobre la base de datos retrospectiva
% "cruda" (sin tratar)

% Cargo los datos de la base retrospectiva:
datos_tabla = readtable('DatosEPOC.xlsx');

% Guardo el orden original de las variables
orden_original = datos_tabla.Properties.VariableNames;

% Elimino variables identificativas, relacionadas con el reingreso (ya
que
% es lo que se va a predecir), fechas y todas aquellas que han sido
% incluidas simplemente para facilitar el proceso de recogida de datos:
variables_elim = {'REV','VALIDO','NH', 'Nombre', 'Apellido_1',
'Apellido_2', ...
    'R_Causa', 'R_Serv', 'F_Ingreso', 'F_Alta', 'F_Nacimiento', ...
    'F_Exitus', 'F_Reingreso', 'R_Exitus','N_Dias_Tran','Dias_alt_exi',
...
```

```
    'R_N_Ur_Hosp', 'R_N_NMLC', 'A_Neumo', 'B_Minhaladora',
'Test_CAT_01','EPOC_Eosinofilico_Y', ...
    'EPOC_Eosinofilico_O', 'I_MLargaDuracion','A_Minhaladora',
'I_Minhaladora','Exitus_30', 'Reingreso'};

datos_tabla(:, variables_elim) = [];

% Además, necesito eliminar aquellos pacientes que durante el ingreso han
% fallecido y que, por tanto, no han podido entrar en el seguimiento de
los
% 30 días post alta para predecir su reingreso. Estos son aquellos
pacientes
% con la variable "A_Exitus = 1":
datos_tabla = datos_tabla(datos_tabla.('A_Exitus') ~= 1, :);
variable_exitus_elim = {'A_Exitus'};

datos_tabla(:, variable_exitus_elim) = [];

% Calculo el porcentaje de NaN en cada columna:
porcentaje_NaN_columnas = sum(ismissing(datos_tabla), 1) /
height(datos_tabla) * 100;
disp('% de NaN por cada variable');
nombres_variables = datos_tabla.Properties.VariableNames;

for i = 1:length(nombres_variables)
    fprintf('%s: %.2f%%\n', nombres_variables{i},
porcentaje_NaN_columnas(i));
end


% Creo una nueva fila cuyos elementos son los porcentajes de NaN de cada
variable
porcentaje_NaN_ant_columnas = zeros(1, length(orden_original));

for i = 1:length(orden_original)

    if ismember(orden_original{i}, datos_tabla.Properties.VariableNames)
        idx = strcmp(datos_tabla.Properties.VariableNames,
orden_original{i});
        porcentaje_NaN_ant_columnas(i) = porcentaje_NaN_columnas(idx);
    else
        porcentaje_NaN_ant_columnas(i) = NaN;
    end
end

% Me quedo solo con las variables finales
variables_finales = datos_tabla.Properties.VariableNames;
idx_finales = ismember(orden_original, variables_finales);
porcentaje_NaN_usar = porcentaje_NaN_ant_columnas(idx_finales);
```

```
fila_nan = array2table(porcentaje_NaN_usar, 'VariableNames',
variables_finales);

% Concateno la fila de porcentajes de datos perdidos por cada variable a
la
% tabla final, siendo esta fila la última en la base de datos del Excel
nueva_Tabla_filaPerdColum = [datos_tabla; fila_nan];


% Calculo el % de NaNs por fila (paciente):
datosPerdidos_filas = sum(ismissing(datos_tabla), 2) / width(datos_tabla)
* 100;
fprintf('\nPorcentaje de NaN por fila:\n');

for i = 1:length(datosPerdidos_filas)
    fprintf('Fila %d: %.2f%%\n', i, datosPerdidos_filas(i) );
end

% Creo una nueva columna con los valores de NaNs por fila
columna_perdidas = num2cell(datosPerdidos_filas);
columna_perdidas = [columna_perdidas; {NaN}];
tabla_perdidas = cell2table(columna_perdidas, 'VariableNames', {'%
perd_filas'});

% Concateno horizontalmente, de forma que la columna de NaNs por fila
ocupe
% la última columna de la base de datos
nueva_Tabla_filaPerdColum = [nueva_Tabla_filaPerdColum, tabla_perdidas];

% Guardo el resultado en Excel y nombro el archivo. La base de datos
% retrospectiva contendrá ahora una última fila de datos perdidos por
% variable y una última columna de NaNs por paciente:
writetable(nueva_Tabla_filaPerdColum,
'DatosEPOC_Porcentajes_NaN_FINAL.xlsx');
```

En segundo lugar, para facilitar la identificación de variables que superasen el umbral de pérdidas impuesto, se elaboró un script que permitiese visualizar de color rojo dichas variables:

```
%----------------------------------------------------------------------
%----------------------------------------------------------------------
%IDENTIFICACIÓN VISUAL DE LAS COLUMNAS (VARIABLES) CUYO PORCENTAJE DE
DATOS
%PERDIDOS ES > 10%.
%----------------------------------------------------------------------
%----------------------------------------------------------------------


nombre_archivo = 'DatosEPOC_Porcentajes_NaN_FINAL.xlsx';
hoja = 1;
umbral = 10; % Establezco un umbral del 10% de NaNs por columna.

% Cargo la tabla conservando el nombre original de las columnas:
elem_tabla = readtable(nombre_archivo, 'VariableNamingRule', 'preserve');
nombres_variables = elem_tabla.Properties.VariableNames;

% Identifico y guardo la última fila del script que es la que contiene el
número de NaNs por columna:
fila_nan = elem_tabla(end, :);

% Se identifican las variables con más del 10% de datos perdidos.
columnas_mas_10_NaN = {}; %Inicializo

for i = 1:width(fila_nan)
    valor = fila_nan{1, i};
    if isnumeric(valor) && ~isnan(valor) && valor > umbral
        columnas_mas_10_NaN{end+1} = nombres_variables{i}; % Recojo los
nombres de las variables que superan el umbral.
    end
end


excel = actxserver('Excel.Application');
excel.Visible = true;
wb = excel.Workbooks.Open(fullfile(pwd, nombre_archivo));
ws = wb.Sheets.Item(hoja);

% Se pinta de rojo las columnas que superan el 10% de NaNs:
for i = 1:length(columnas_mas_10_NaN)
    nombre_col = columnas_mas_10_NaN{i};
    fprintf("Columna con > %d%% NaN: %s\n", umbral, nombre_col);
    col_index = find(strcmp(nombres_variables, nombre_col));

    if ~isempty(col_index)
        col_letra = obtener_letra_col_excel(col_index);
```

```
        rango_col = sprintf('%s2:%s%d', col_letra, col_letra,
height(elem_tabla) + 1); % Se añade "+1" porque la fila que contiene el
porcentaje de NaNs por columna es la última de todas.
        ws.Range(rango_col).Interior.Color = 255; % 255 es el número
correspondiente al color rojo.
    end
end

% Función auxiliar
function letra = obtener_letra_col_excel(n)
% Transforma un número de columna a su letra correspondiente en Excel, ya
% que este último identifica las columnas por letras.
%Entrada: n - Valor numérico del índice de la columna.
%Salida: letra - String de la letra que se corresponde a la columna de
%Excel.

    % Convierte un número de columna a letra estilo Excel (1 -> A, 27 ->
AA)
    letra = '';
    while n > 0
        resto_div = mod(n - 1, 26);
        letra = [char(65 + resto_div) letra];
        n = floor((n - 1) / 26);
    end
end
```

Tras esto, se muestra el código mediante el cual se prescinde de las variables que superan el umbral y de aquellas cuyo sumatorio de valores resultase ser 0 o 1. Tras esta eliminación, se presenta el cálculo de nuevo de los datos perdidos, tanto por paciente como por variable, comprobado que el tratamiento de estos datos se ha realizado correctamente:

```
%-----------------------------------------------------------------------
%-----------------------------------------------------------------------
% FILTRADO DE LA BASE DE DATOS RETROSPECTIVA
%-----------------------------------------------------------------------
%-----------------------------------------------------------------------

% Cargo la tabla que contiene el porcentaje de los datos perdidos.
datos_tabla = readtable('DatosEPOC_Porcentajes_NaN_FINAL.xlsx',
'VariableNamingRule', 'preserve');

% Separar la última fila (porcentajes de NaNs por columna)
% Identifico y separo la última columna, que es la que contiene el
porcentaje de NaNs.

fila_nan = datos_tabla(end, :); % Selecciono y recojo en una variable la
última fila (% NaNs).
datos_tabla_final = datos_tabla(1:end-1, :);   % Selecciono y recojo en
una variable el resto de pacientes


% Se establece el umbral de datos perdidos, eliminado aquellas variables
% que superen el mismo:
umbral = 10;
columnas_mas_10_NaN = {}; % Inicializo
variable_names = datos_tabla.Properties.VariableNames;

for i = 1:width(fila_nan)
    valor = fila_nan{1, i};
    if isnumeric(valor) && ~isnan(valor) && valor > umbral
        columnas_mas_10_NaN{end+1} = variable_names{i};
    end
end

fprintf('Columnas que presentan más de un 10% de NaNs:\n');
disp(columnas_mas_10_NaN); % Muestro las columnas que han superado el
umbral.

datos_filtrados = removevars(datos_tabla_final, columnas_mas_10_NaN);

% A continuación, se eliminan aquellas variables cuyo sumatorio es 0 o 1
% (no tienen suficiente varianza como para discriminar entre clases):

num_columnas = width(datos_filtrados); % Calculo el número de variables
```

```
columnas_numericas = false(1, num_columnas); %  Se identifican las
columnas numéricas.

for i = 1:num_columnas
    columnas_numericas(i) = isnumeric(datos_filtrados{:, i}); % Se
comprueba si cada columna es o no numérica.
end

datos_numericos = datos_filtrados(:, columnas_numericas);


% Se obtiene el sumatorio de los elementos de cada variable omitiendo los
% NaNs:
num_columnas_numericas = width(datos_numericos);
result_sum_colum = zeros(1, num_columnas_numericas); % Inicializo la
variable donde se guardarán los sumatorios.

for i = 1:num_columnas_numericas
    result_sum_colum(i) = sum(datos_numericos{:, i}, 'omitnan'); %
Sumatorio por cada variable.
end

columnas_a_eliminar = result_sum_colum == 0 | result_sum_colum == 1; %
Aplico la condición: suma 0 o 1.

nombres_columnas_eliminar =
datos_numericos.Properties.VariableNames(columnas_a_eliminar);
datos_filtrados = removevars(datos_filtrados, nombres_columnas_eliminar);
% Elimino dichas variables.

% Elimino también la última columna, que contiene el porcentaje de NaNs
por
% fila, dado que al eliminar variables, los porcentajes de datos perdidos
% por paciente han variado y ya no se corresponden con los previos.
datos_filtrados(:, end) = [];


% Calculo de nuevo el porcentaje de NaNs por paciente tras el filtrado de
% variables
porcentaje_nan_filas = sum(ismissing(datos_filtrados), 2) /
width(datos_filtrados) * 100;

% Añado esa columna a la tabla de datos.
datos_filtrados.Porcentaje_NaNs_Fila = porcentaje_nan_filas;


% Tras el filtrado de variables, se calculan de nuevo los porcentajes de
% datos perdidos por columnas que, aunque no varían, sirve para comprobar
% que ninguna de las variables restantes contiene más de un 10% de NaNs.
```

```
% Además, estas características serán las proporcionadas más tarde al
% algoritmo de selección de variables.
porcentaje_fila_final = sum(ismissing(datos_filtrados), 1) /
height(datos_filtrados) * 100;
fila_porcentaje_columnas = array2table(porcentaje_fila_final,
'VariableNames', datos_filtrados.Properties.VariableNames);

% Añado la fila de datos perdidos por variable como última fila.
datos_completo = [datos_filtrados; fila_porcentaje_columnas];


% Se muestran los resultados:
fprintf('\nPorcentaje de NaNs por paciente después del filtrado:\n');

for i = 1:length(porcentaje_nan_filas)
    fprintf('Fila %d: %.2f%%\n', i, porcentaje_nan_filas(i));
end

fprintf('\nPorcentaje de NaNs por variable después del filtrado:\n');

for j = 1:width(datos_filtrados)
    fprintf('%s: %.2f%%\n', datos_filtrados.Properties.VariableNames{j},
porcentaje_fila_final(j));
end


% Se guarda el resultado en un Excel nuevo:
writetable(datos_completo, 'DatosEPOC_Filtrados_Completo_FINAL.xlsx');
```

A continuación, se expone el *script* elaborado para la imputación (*K* vecinos más cercanos), con un número de vecinos de 5. En las variables categóricas, se emplea la moda (voto mayoritario o *majority rule*), mientras que para las continuas se aplica la media (*mean rule*).

```
%--------------------------------------------------------------------------
%--------------------------------------------------------------------------
% IMPUTACIÓN DE DATOS: K VECINOS MÁS CERCANOS
%--------------------------------------------------------------------------
%--------------------------------------------------------------------------

% Se cargan los datos ya filtrados (con las eliminaciones de columnas
% pertinentes):
datos_vecinos = readtable("DatosEPOC_Filtrados_Completo_FINAL.xlsx");

datos_tabla_origen = readtable('DatosEPOC.xlsx');
datos_tabla_origen = datos_tabla_origen(datos_tabla_origen.('A_Exitus')
~= 1, :);
variable_reingreso = {'Reingreso'}
datos_reingreso = datos_tabla_origen{:, variable_reingreso}

variables_continuas = {'Edad', 'Peso', 'Altura', 'IMC',
'I_Tabaquico','Test_Charlson', ...
    'B_Esp_FVC', 'B_Esp_FVC_p', 'B_Esp_FVC_iz', 'B_Esp_FEV1',
'B_Esp_FEV1_p', ...
    'B_Esp_FEV1_iz', 'B_Esp_FEV1FVC', 'B_Esp_FEV1FVC_p',
'B_Esp_FEV1FVC_iz', ...
    'N_Ingresos', 'Test_Barthel','Test_CAT',
'N_Dias_Ing','N_Dias_Sintomas', 'I_CV_TAS', 'I_CV_TAD',...
    'I_CV_FC', 'I_Gas_PH', 'I_Gas_PCO2', 'I_Gas_PO2', 'I_Gas_HCO3',...
    'I_Ana_Leuc', 'I_Ana_Neu', 'I_Ana_Neup','I_Ana_Eos', 'I_Ana_Eosp'};

% Variablescategóricas:
variables_categoricas = {'Sexo', 'Procedencia', 'E_Civil', 'Hogar',
'Estudios', 'A_Laboral', ...
    'Movilidad', 'Cuidador', 'Tabaquismo','Alcohol', 'GRC', 'Anti_Coag',
'Anti_Agreg', ...
    'V_Gripe', 'V_Neumo', 'HTA', 'DM', 'Dislipemia', 'Card_Isq',
'Ins_Card', 'Ins_Resp', ...
    'Bronqui', 'Taquiarritmia', 'ACV', 'Neo_Pulmon', 'Neo_Otras', ...
    'Enf_Renal', 'Osteopor', 'Ansiedad', 'Depresion', 'Anemia', 'TEP',
'SAHS', 'Estr_Riesgo', ...
    'Fenotipo', 'Grd_Obst', 'Gold', 'B_Oxi_Dom', 'B_VNI',
'B_Min_Ninguna', 'B_Min_SABA', ...
    'B_Min_SAMA', 'B_Min_LABA', 'B_Min_LAMA', 'B_Min_CI',
'B_Mre_Ninguna', 'B_Mre_SABA', ...
    'B_Mre_SAMA', 'B_Mre_LABA', 'B_Mre_LAMA', 'B_Mre_CI', 'B_Corti',
'B_Teo', 'B_IDFE4', ...
    'B_Muco', 'B_Antibio', 'B_FisioRes','Test_mMRC','Test_TAI_Ade', ...
```

```
    'Test_TAI_I_Err', 'Test_TAI_I_Del', 'Test_TAI_I_Inc', 'Test_E5D_Mov',
'Test_E5D_CPe', ...
    'Test_E5D_ACo', 'Test_E5D_Dol', 'Test_E5D_Dep', ...
    'Aumento_Tos', 'Aumento_Disnea', 'Aumento_Expect', 'Purul_Esputo',
'Dolor_Tracico', 'Fiebre', ...
    'Uso_Muscul_Acce', 'Mov_Toracicos', 'Cianosis', 'Edemas_Per',
'Inest_Hemod', 'Deter_Mental', ...
    'Disnea', 'Causa_Infecciosa', 'Causa_Bacteriana', 'Causa_Virica',
'Microorg_Resis', ...
    'Comp_Arritmias', 'Comp_Insuf_Cardia', 'Comp_Card_Isquem',
'Comp_Derr_Pleural', ...
    'Comp_Neumonia', 'Comp_Sepsis', 'Comp_Insuf_Respir', ...
    'I_Min_Ninguna', 'I_Min_SABA', 'I_Min_SAMA', 'I_Min_LABA',
'I_Min_LAMA', 'I_Min_CI', ...
    'I_Corti', 'I_UVI', 'I_Oxi', 'I_Teo', 'I_IDFE4', ...
    'I_Muco', 'I_VNI', 'I_VIN', 'I_Antibio', 'A_Oxi_Dom', 'A_VNI',
'A_Min_Ninguna', ...
    'A_Min_SABA', 'A_Min_SAMA', 'A_Min_LABA', 'A_Min_LAMA', 'A_Min_CI',
...
    'A_Mre_Ninguna', 'A_Mre_SABA', 'A_Mre_SAMA', 'A_Mre_LABA',
'A_Mre_LAMA', 'A_Mre_CI', ...
    'A_Corti', 'A_Antibio', 'A_Teo', 'A_IDFE4', 'A_Muco', 'A_FisioRes'};


if height(datos_vecinos) > length(datos_reingreso)
    datos_reingreso = [datos_reingreso; 0];
end
datos_vecinos.Reingreso = datos_reingreso;

indice_porcentaje_nan =
find(strcmp(datos_vecinos.Properties.VariableNames,
'Porcentaje_NaNs_Fila'));
datos_vecinos(:, indice_porcentaje_nan) = [];


columna_reingreso = find(strcmp(datos_vecinos.Properties.VariableNames,
'Reingreso')); % Se guarda el índice correspondiente a la variable target
(reingreso)
datos_matriz = datos_vecinos{:, :}; %Guardamos como matriz

% Dado que la última fila y columna se correspondían con el porcentaje de
NaNs por variable y paciente respectivamente, se eliminan para la
imputación:
datos_matriz(end, :) = []; % Se elimina la fila de NaNs por variable
(última fila de la base de datos).

% datos_matriz(:, end) = []; % Se elimina la última columna de NaNs por
paciente (última columna de la base de datos).
```

```
reingreso_col = datos_matriz(:, columna_reingreso); % Se guarda en una
variable la columna correspondiente a la variable target "reingreso".

% Se excluye la variable target (reingreso), ya que esta no debe
intervenir
% en la etapa de imputación de datos perdidos:
datos_matriz(:, columna_reingreso) = [];

% Dado que la base de datos posee variables de carácter mixto, se
discrimina
% entre variables continuas y categóricas para la imputación.

%En primer lugar, se considera categórica las variables con valores sin
%parte decimal. Para ello, se aplica una tolerancia, determinando si se
%trata o no de un valor entero:
num_columnas = size(datos_matriz, 2);
esCategorica = false(1, num_columnas);
variable_names = datos_vecinos.Properties.VariableNames(setdiff(1:end,
columna_reingreso));
for j = 1:num_columnas
    esCategorica(j) = ismember(variable_names{j}, variables_categoricas);
end


[num_nan, ~] = size(datos_matriz);
[filas, colum] = find(isnan(datos_matriz)); % Se recogen los índices que
contienen los NaNs.

k = 5; % Número de vecinos implementado.

for i = 1:length(filas)

    fila_idx = filas(i);
    col_idx = colum(i);

    % Se extrae la fila actual y se elimina la columna del valor perdido.
    Y = datos_matriz(fila_idx, :);
    Y(:, col_idx) = [];

    % Guardar la clase a la que pertenece el paciente de la fila actual:
    clase_reingreso = reingreso_col(fila_idx);

    % De X, se conserva únicamente los pacientes del mismo grupo o clase:
    X = datos_matriz(reingreso_col == clase_reingreso, :);
    X(:, col_idx) = [];

    % Se aplica K vecinos más cercanos mediante la función knnsearch:
    idx = knnsearch(X, Y, 'K', k);
```

```
    % Mapear los índices de X a los índices originales en datos_matriz
    idx_original = find(reingreso_col == clase_reingreso);
    idx_original = idx_original(idx);

    % Para la imputación, en las variables continuas se calcula la media
de
    % los K vecinos más cercanos, mientras que para las categóricas se
    % emplea la moda:
    if esCategorica(col_idx)
        % Variables categóricas: moda (voto mayoritario o majority rule)
        imputado = mode(datos_matriz(idx_original, col_idx));
    else
        % Variables continuas: media (mean rule)
        imputado = mean(datos_matriz(idx_original, col_idx), 'omitnan');
    end

    % El valor imputado calculado se sustituye en la posición que
corresponda:
    datos_matriz(fila_idx, col_idx) = imputado;
end


% Se elimina la última fila y columna de la tabla original para que sus
% dimensiones coincidan con datos_matriz:
datos_vecinos(end, :) = [];

datos_vecinos(:, end) = [];

% Se actualiza la base de datos los resultados de la imputación, sin
% incluir la variable target:
datos_vecinos{:, setdiff(1:size(datos_vecinos, 2), columna_reingreso)} =
datos_matriz;

% Se guardan los resultados en un archivo Excel:
archivo_kvecinos = 'Datos_imputacionkVecinosAplicado_FINAL.xlsx';
writetable(datos_vecinos, archivo_kvecinos);
```

ANEXO 2. Código análisis descriptivo de variables.

En este código, se incluye el cálculo de estadísticos descriptivos de las variables que forman parte de las bases de datos, tanto de la retrospectiva como la prospectiva. Se distingue entre variables continuas y categóricas. Para las primeras, se calcula la mediana y primer y tercer cuartil, mientras que para las segundas se efectúa un recuento de pacientes por cada categoría y se expresa en porcentaje. Todos los estadísticos calculados mediante el siguiente código se recogen en una tabla conformada por 4 columnas. La primera columna se corresponde con el grupo total de pacientes, la segunda con la clase negativa (no reingreso), la tercera con la clase positiva (reingreso) y finalmente, la cuarta con el p-valor. Este último se obtiene mediante el Test de Fisher para las variables categóricas y el Test de Mann-Whitney para las continuas.

Cohorte retrospectiva

```
% Cargo la base de datos:
datos_tabla = readtable('Datos_imputacionkVecinosAplicado_FINAL.xlsx',
'VariableNamingRule', 'preserve');

datos_tabla_origen = readtable('DatosEPOC.xlsx');
datos_tabla_origen = datos_tabla_origen(datos_tabla_origen.('A_Exitus')
~= 1, :);
variable_reingreso = {'Reingreso'}
datos_reingreso = datos_tabla_origen{:, variable_reingreso}

% Defino las variables de interés:
% Variables continuas:
variables_continuas = {'Edad', 'Peso', 'Altura', 'IMC',
'I_Tabaquico','Test_Charlson', ...
    'B_Esp_FVC', 'B_Esp_FVC_p', 'B_Esp_FVC_iz', 'B_Esp_FEV1',
'B_Esp_FEV1_p', ...
    'B_Esp_FEV1_iz', 'B_Esp_FEV1FVC', 'B_Esp_FEV1FVC_p',
'B_Esp_FEV1FVC_iz', ...
    'N_Ingresos', 'Test_Barthel','Test_CAT',
'N_Dias_Ing','N_Dias_Sintomas', 'I_CV_TAS', 'I_CV_TAD',...
    'I_CV_FC', 'I_Gas_PH', 'I_Gas_PCO2', 'I_Gas_PO2', 'I_Gas_HCO3',...
    'I_Ana_Leuc', 'I_Ana_Neu', 'I_Ana_Neup','I_Ana_Eos', 'I_Ana_Eosp'};

% Variablescategóricas:
variables_categoricas = {'Sexo', 'Procedencia', 'E_Civil', 'Hogar',
'Estudios', 'A_Laboral', ...
    'Movilidad', 'Cuidador', 'Tabaquismo','Alcohol', 'GRC', 'Anti_Coag',
'Anti_Agreg', ...
    'V_Gripe', 'V_Neumo', 'HTA', 'DM', 'Dislipemia', 'Card_Isq',
'Ins_Card', 'Ins_Resp', ...
    'Bronqui', 'Taquiarritmia', 'ACV', 'Neo_Pulmon', 'Neo_Otras', ...
    'Enf_Renal', 'Osteopor', 'Ansiedad', 'Depresion', 'Anemia', 'TEP',
'SAHS', 'Estr_Riesgo', ...
```

```
    'Fenotipo', 'Grd_Obst', 'Gold', 'B_Oxi_Dom', 'B_VNI',
'B_Min_Ninguna', 'B_Min_SABA', ...
    'B_Min_SAMA', 'B_Min_LABA', 'B_Min_LAMA', 'B_Min_CI',
'B_Mre_Ninguna', 'B_Mre_SABA', ...
    'B_Mre_SAMA', 'B_Mre_LABA', 'B_Mre_LAMA', 'B_Mre_CI', 'B_Corti',
'B_Teo', 'B_IDFE4', ...
    'B_Muco', 'B_Antibio', 'B_FisioRes','Test_mMRC','Test_TAI_Ade', ...
    'Test_TAI_I_Err', 'Test_TAI_I_Del', 'Test_TAI_I_Inc', 'Test_E5D_Mov',
'Test_E5D_CPe', ...
    'Test_E5D_ACo', 'Test_E5D_Dol', 'Test_E5D_Dep', ...
    'Aumento_Tos', 'Aumento_Disnea', 'Aumento_Expect', 'Purul_Esputo',
'Dolor_Tracico', 'Fiebre', ...
    'Uso_Muscul_Acce', 'Mov_Toracicos', 'Cianosis', 'Edemas_Per',
'Inest_Hemod', 'Deter_Mental', ...
    'Disnea', 'Causa_Infecciosa', 'Causa_Bacteriana', 'Causa_Virica',
'Microorg_Resis', ...
    'Comp_Arritmias', 'Comp_Insuf_Cardia', 'Comp_Card_Isquem',
'Comp_Derr_Pleural', ...
    'Comp_Neumonia', 'Comp_Sepsis', 'Comp_Insuf_Respir', ...
    'I_Min_Ninguna', 'I_Min_SABA', 'I_Min_SAMA', 'I_Min_LABA',
'I_Min_LAMA', 'I_Min_CI', ...
    'I_Corti', 'I_UVI', 'I_Oxi', 'I_Teo', 'I_IDFE4', ...
    'I_Muco', 'I_VNI', 'I_VIN', 'I_Antibio', 'A_Oxi_Dom', 'A_VNI',
'A_Min_Ninguna', ...
    'A_Min_SABA', 'A_Min_SAMA', 'A_Min_LABA', 'A_Min_LAMA', 'A_Min_CI',
...
    'A_Mre_Ninguna', 'A_Mre_SABA', 'A_Mre_SAMA', 'A_Mre_LABA',
'A_Mre_LAMA', 'A_Mre_CI', ...
    'A_Corti', 'A_Antibio', 'A_Teo', 'A_IDFE4', 'A_Muco', 'A_FisioRes'};


datos_tabla.Reingreso = datos_reingreso;

grupo_no_reingreso = datos_tabla.Reingreso == 0; % Almaceno el grupo de
no reingreso.
grupo_reingreso = datos_tabla.Reingreso == 1; % Almaceno el grupo
reingreso.

% Calculo el número de pacientes de cada clase:
fprintf('Tamaño grupo "no reingreso": %d\n', sum(grupo_no_reingreso));
fprintf('Tamaño grupo "reingreso": %d\n', sum(grupo_reingreso));

% Inicializo la tabla de resultados:
resultados = table();

% Calculo los estadísticos descriptivos para las variables previas:
for i = 1:length(variables_continuas)
    variable_actual = variables_continuas{i};
```

```
    datos_no_reingreso = datos_tabla{grupo_no_reingreso,
variable_actual};
    datos_reingreso = datos_tabla{grupo_reingreso, variable_actual};

    datos_total = datos_tabla{:, variable_actual};

    % Me quedo con los que no contegan NaN:
    valid_no_reingreso = sum(~isnan(datos_no_reingreso));
    valid_reingreso = sum(~isnan(datos_reingreso));
    valid_total = sum(~isnan(datos_total));

    fprintf('Variable: %s, Datos válidos "no reingreso" : %d,
"reingreso": %d, Total: %d\n', ...
        variable_actual, valid_no_reingreso, valid_reingreso,
valid_total);

    % Calculo la mediana, primer y tercer cuartil:
    mediana_no_reingreso = NaN;
    q1_no_reingreso = NaN;
    q3_no_reingreso = NaN;

    mediana_reingreso = NaN;
    q1_reingreso = NaN;
    q3_reingreso = NaN;

    mediana_total = NaN;
    q1_total = NaN;
    q3_total = NaN;

    p_valor = NaN;

    if valid_no_reingreso > 0
        mediana_no_reingreso = median(datos_no_reingreso, 'omitnan');

        q1_no_reingreso = quantile(datos_no_reingreso, 0.25); % Primer
cuartil.
        q3_no_reingreso = quantile(datos_no_reingreso, 0.75); % Tercer
cuartil.
    end

    if valid_reingreso > 0
        mediana_reingreso = median(datos_reingreso, 'omitnan');

        q1_reingreso = quantile(datos_reingreso, 0.25); % Primer cuartil.
        q3_reingreso = quantile(datos_reingreso, 0.75); % Tercer cuartil.
    end

    if valid_total > 0
        mediana_total = median(datos_total, 'omitnan');
```

```
        q1_total = quantile(datos_total, 0.25); % Primer cuartil.
        q3_total = quantile(datos_total, 0.75); % Tercer cuartil.
    end

    if valid_no_reingreso >= 1 && valid_reingreso >= 1
        try
            [p_valor, ~]= ranksum(datos_no_reingreso, datos_reingreso); %
Test Mann-Whitney
        catch
            fprintf('Imposible aplicar Mann-Whitney para %s: faltan
datos\n', variable_actual);
        end
    end

    % Genero la tabla de estadísticos descriptivos:
    resultados =[resultados; table({variable_actual}, ...
        {[num2str(mediana_total, '%.1f') ' (' num2str(q1_total, '%.1f')
', ' num2str(q3_total, '%.1f') ')']}, ...
        {[num2str(mediana_no_reingreso, '%.1f') ' ('
num2str(q1_no_reingreso, '%.1f') ', ' num2str(q3_no_reingreso, '%.1f')
')']}, ...
        {[num2str(mediana_reingreso, '%.1f') ' (' num2str(q1_reingreso,
'%.1f') ', ' num2str(q3_reingreso, '%.1f') ')']}, ...
        {p_valor}, 'VariableNames', {'Variable', 'Total', 'No reingreso',
'Reingreso', 'p_valor'})];
end


% Calculo en forma de porcentaje los pacientes que hay en cada categoría:
for i = 1:length(variables_categoricas)
    variable = variables_categoricas{i};
    if ~iscategorical(datos_tabla.(variable))
        datos_tabla.(variable) = categorical(datos_tabla.(variable));
    end

    % Por cada categoría, calculo ese porcentaje:
    categorias = categories(datos_tabla.(variable));

    for j = 1:length(categorias)

        categoria = categorias{j};
        fprintf(variable + " "+ categoria +"\n");

        if ~(strcmp(variable, 'Test_E5D_Dol') && strcmp(categoria, '0'))

            sum_no_reingreso = sum(datos_tabla{grupo_no_reingreso,
variable} == categoria);
```

```
            sum_reingreso = sum(datos_tabla{grupo_reingreso, variable} ==
categoria);

            total_no_reingreso = sum(grupo_no_reingreso);
            total_reingreso = sum(grupo_reingreso);
            total_sample =height(datos_tabla);
            sum_total = sum(datos_tabla{:, variable} == categoria);

            porcentaje_no_reingreso = (sum_no_reingreso /
total_no_reingreso) * 100;
            porcentaje_reingreso = (sum_reingreso / total_reingreso) *
100;
            porcentaje_total = (sum_total / total_sample) * 100;

            tabla_contingencia =[sum_no_reingreso, total_no_reingreso -
sum_no_reingreso; ...
                                 sum_reingreso, total_reingreso -
sum_reingreso];

            p_valor = NaN;

            if any(tabla_contingencia(:) > 0) % Manejo de posibles datos
faltantes.
                try
                    [~, p_valor] = fishertest(tabla_contingencia); %
Aplico el Test de Fisher.
                catch
                    fprintf('Imposible aplicar el Test de Fisher para %s
(%s): faltan datos\n', variable, categoria);
                end
            end

            % Muestro los resultados del test en la tabla:
            resultados =[resultados; table({[variable ' (' categoria
')']}, ...
                {[num2str(sum_total) ' (' num2str(porcentaje_total,
'%.1f') '%)']}, ...
                {[num2str(sum_no_reingreso) ' ('
num2str(porcentaje_no_reingreso, '%.1f') '%)']}, ...
                {[num2str(sum_reingreso) ' ('
num2str(porcentaje_reingreso, '%.1f') '%)']}, ...
                {p_valor}, ...
                'VariableNames', {'Variable', 'Total', 'No reingreso',
'Reingreso', 'p_valor'})];

        end
    end
end
```

```
% Muestro los resultados de los estadísticos, tanto de variables
continuas como categóricas, en tabla:
disp(resultados);

% Guardo los resultados en un nuevo Excel:
writetable(resultados, 'Resultados_EstDescriptivos_postNan.xlsx');


% Filtrar variables continuas con p-valor < 0.05
significativas = {};
for i = 1:height(resultados)
    var = resultados.Variable{i};
    if ismember(var, variables_continuas) &&
~isnan(resultados.p_valor{i}) && resultados.p_valor{i} < 0.05
        significativas = [significativas; var];
    end
end

num_sig = length(significativas);

if num_sig == 0
    disp('No hay variables continuas con diferencias significativas.');
else
    % Calcular cuántas figuras se necesitan (máximo 2 subplots por
figura)
    num_figs = ceil(num_sig);
    for f = 1:num_figs
        figure;
        for s = 1
            idx = (f-1)+ s;
            if idx > num_sig
                break;
            end
            var = significativas{idx};

            % Preparar datos para el boxplot
            datos_no = datos_tabla{grupo_no_reingreso, var};
            datos_re = datos_tabla{grupo_reingreso, var};
            data = [datos_no; datos_re];
            group = [ones(length(datos_no).1);
2*ones(length(datos_re).1)];

            % Crear subplot y boxplot
            boxplot(data, group, 'Notch', 'off', 'Labels', {'No
Reingreso', 'Reingreso'});
            title(['Diagrama de cajas para la variable ' var],
'FontSize', 12, 'Interpreter', 'none');
            ylabel(['Valores de la variable ' var], 'FontSize', 12,
'Interpreter', 'none');
```

```
            xlabel('Grupos bajo estudio', 'FontSize', 12);

            if ~exist('C:\plot', 'dir')
                mkdir('C:\plot');
            end
            print(['C:\\plot\\graf' num2str(f) '.png'], '-dpng');
        end
    end
end
```

A continuación, se exhibe también el análisis descriptivo de las variables relacionadas directamente con el reingreso. Su estudio estadístico se realiza por separado, ya que estas variables no deben ser filtradas atendiendo a su porcentaje de datos perdidos, ni deben formar parte del conjunto candidato de variables de entrada a los modelos predictivos. Su valor en este trabajo es meramente informativo.

```
%Cargamos la base de datos original y recogemos las variables especiales
datos_tabla = readtable('DatosEPOC.xlsx');
% Eliminamos pacientes con A_Exitus = 1
datos_tabla = datos_tabla(datos_tabla.('A_Exitus') ~= 1, :);

pacientes_reingreso = datos_tabla(datos_tabla.('Reingreso') ~= 0, :);
porcentaje_r_exitus = mean(pacientes_reingreso.R_Exitus)*100;
media_n_dias_tran = mean(pacientes_reingreso.N_Dias_Tran);

fprintf('porcentaje_r_exitus = %.2f%%\n', porcentaje_r_exitus);
fprintf('media_n_dias_tran = %.2f dias\n', media_n_dias_tran);

variables_guardar = { 'Dias_alt_exi','R_N_Ur_Hosp', 'R_N_NMLC',
'Exitus_30', 'Reingreso'};
tabla_filtrada = datos_tabla(:, variables_guardar);

% Guardar la tabla filtrada en un nuevo archivo Excel
writetable(tabla_filtrada, 'DatosEPOC_especiales.xlsx');
% Cargo la base de datos:

datos_tabla = readtable('DatosEPOC_especiales.xlsx',
'VariableNamingRule', 'preserve');
% Defino las variables de interés:
% Variables continuas:
variables_continuas = {'Dias_alt_exi' };
% Variablescategóricas:
variables_categoricas = {'R_N_Ur_Hosp', 'R_N_NMLC', 'Exitus_30'};
grupo_no_reingreso = datos_tabla.Reingreso == 0; % Almaceno el grupo de
no reingreso.
```

```
grupo_reingreso = datos_tabla.Reingreso == 1; % Almaceno el grupo
reingreso.
% Calculo el número de pacientes de cada clase:
fprintf('Tamaño grupo "no reingreso": %d\n', sum(grupo_no_reingreso));
fprintf('Tamaño grupo "reingreso": %d\n', sum(grupo_reingreso));

% Inicializo la tabla de resultados:
resultados = table();
% Calculo los estadísticos descriptivos para las variables previas:
for i = 1:length(variables_continuas)
    variable_actual = variables_continuas{i};
    datos_no_reingreso = datos_tabla{grupo_no_reingreso,
variable_actual};
    datos_reingreso = datos_tabla{grupo_reingreso, variable_actual};
    datos_total = datos_tabla{:, variable_actual};

    % Me quedo con los que no contegan NaN:
    valid_no_reingreso = sum(~isnan(datos_no_reingreso));
    valid_reingreso = sum(~isnan(datos_reingreso));
    valid_total = sum(~isnan(datos_total));
    fprintf('Variable: %s, Datos válidos "no reingreso" : %d,
"reingreso": %d, Total: %d\n', ...
        variable_actual, valid_no_reingreso, valid_reingreso,
valid_total);

    % Calculo la mediana, primer y tercer cuartil:
    mediana_no_reingreso = NaN;
    q1_no_reingreso = NaN;
    q3_no_reingreso = NaN;

    mediana_reingreso = NaN;
    q1_reingreso = NaN;
    q3_reingreso = NaN;

    mediana_total = NaN;
    q1_total = NaN;
    q3_total = NaN;

    p_valor = NaN;

    if valid_no_reingreso > 0
        mediana_no_reingreso = median(datos_no_reingreso, 'omitnan');
        q1_no_reingreso = quantile(datos_no_reingreso, 0.25); % Primer
cuartil.
        q3_no_reingreso = quantile(datos_no_reingreso, 0.75); % Tercer
cuartil.
    end
    if valid_reingreso > 0
        mediana_reingreso = median(datos_reingreso, 'omitnan');
```

```
        q1_reingreso = quantile(datos_reingreso, 0.25); % Primer cuartil.
        q3_reingreso = quantile(datos_reingreso, 0.75); % Tercer cuartil.
    end
    if valid_total > 0
        mediana_total = median(datos_total, 'omitnan');
        q1_total = quantile(datos_total, 0.25); % Primer cuartil.
        q3_total = quantile(datos_total, 0.75); % Tercer cuartil.
    end

    if valid_no_reingreso >= 1 && valid_reingreso >= 1
        try
            [p_valor, ~]= ranksum(datos_no_reingreso, datos_reingreso); %
Test Mann-Whitney
        catch
            fprintf('Imposible aplicar Mann-Whitney para %s: faltan
datos\n', variable_actual);
        end
    end

    % Genero la tabla de estadísticos descriptivos:
    resultados =[resultados; table({variable_actual}, ...
        {[num2str(mediana_total, '%.1f') ' (' num2str(q1_total, '%.1f')
', ' num2str(q3_total, '%.1f') ')']}, ...
        {[num2str(mediana_no_reingreso, '%.1f') ' ('
num2str(q1_no_reingreso, '%.1f') ', ' num2str(q3_no_reingreso, '%.1f')
')']}, ...
        {[num2str(mediana_reingreso, '%.1f') ' (' num2str(q1_reingreso,
'%.1f') ', ' num2str(q3_reingreso, '%.1f') ')']}, ...
        {p_valor}, 'VariableNames', {'Variable', 'Total', 'No reingreso',
'Reingreso', 'p_valor'})];
end
% Calculo los estadísticos descriptivos para las variables previas:%
Calculo en forma de porcentaje los pacientes que hay en cada categoría:
for i = 1:length(variables_categoricas)
    variable = variables_categoricas{i};
    if ~iscategorical(datos_tabla.(variable))
        datos_tabla.(variable) = categorical(datos_tabla.(variable),
'Ordinal', false);
    end

    % Obtenenemos las categorías válidas.
    categorias = categories(datos_tabla.(variable));

    for j = 1:length(categorias)
        categoria = categorias{j};

        % Filtro todos los datos no nulos.
        datos_no_reingreso = datos_tabla{grupo_no_reingreso, variable};
        datos_reingreso = datos_tabla{grupo_reingreso, variable};
```

```
        datos_total = datos_tabla{:, variable};

        % Contabilizamos solo valores no nulos.
        sum_no_reingreso = sum(datos_no_reingreso == categoria &
~ismissing(datos_no_reingreso));
        sum_reingreso = sum(datos_reingreso == categoria &
~ismissing(datos_reingreso));
        sum_total = sum(datos_total == categoria &
~ismissing(datos_total));

        total_no_reingreso = sum(~ismissing(datos_no_reingreso));
        total_reingreso = sum(~ismissing(datos_reingreso));
        total_sample = sum(~ismissing(datos_total));

        % Calculaculamos los porcentajes.
        porcentaje_no_reingreso = (sum_no_reingreso / total_no_reingreso)
* 100;
        porcentaje_reingreso = (sum_reingreso / total_reingreso) * 100;
        porcentaje_total = (sum_total / total_sample) * 100;

        % Creamos la tabla de contingencia.
        tabla_contingencia = [sum_no_reingreso, total_no_reingreso -
sum_no_reingreso; ...
                              sum_reingreso, total_reingreso -
sum_reingreso];

        p_valor = NaN;

        % Test de Fisher solo si la tabla tiene datos válidos y no es
singular.
        if all(tabla_contingencia(:) >= 0) && sum(tabla_contingencia(:))
> 0 && ~any(all(tabla_contingencia == 0, 1)) &&
~any(all(tabla_contingencia == 0, 2))
            try
                [~, p_valor] = fishertest(tabla_contingencia);
            catch
                fprintf('Imposible aplicar el Test de Fisher para %s
(%s): datos insuficientes\n', variable, categoria);
            end
        else
            fprintf('Test de Fisher no aplicable para %s (%s): tabla de
contingencia singular\n', variable, categoria);
        end

        % Agregamos a tabla de resultados.
        resultados = [resultados; table({[variable ' (' categoria ')']},
...
            {[num2str(sum_total) ' (' num2str(porcentaje_total, '%.1f')
'%)']}, ...
```

```
            {[num2str(sum_no_reingreso) ' ('
num2str(porcentaje_no_reingreso, '%.1f') '%)']}, ...
            {[num2str(sum_reingreso) ' (' num2str(porcentaje_reingreso,
'%.1f') '%)']}, ...
            {p_valor}, ...
            'VariableNames', {'Variable', 'Total', 'No reingreso',
'Reingreso', 'p_valor'})];
    end
end
% Muestro los resultados de los estadísticos, tanto de variables
continuas como categóricas, en tabla:
disp(resultados);
% Guardo los resultados en un nuevo Excel:
writetable(resultados, 'Resultados_EstDescriptivos_Especiales.xlsx');
% Filtrar variables continuas con p-valor < 0.05.
significativas = {};
for i = 1:height(resultados)
    var = resultados.Variable{i};
    if ismember(var, variables_continuas) &&
~isnan(resultados.p_valor{i}) && resultados.p_valor{i} < 0.05
        significativas = [significativas; var];
    end
end
num_sig = length(significativas);
if num_sig == 0
    disp('No hay variables continuas con diferencias significativas.');
else
    % Calculo cuántas figuras se necesitan.
    num_figs = ceil(num_sig);
    for f = 1:num_figs
        figure;
        for s = 1
            idx = (f-1)+ s;
            if idx > num_sig
                break;
            end
            var = significativas{idx};
            % Diagrama de cajas.
            datos_no = datos_tabla{grupo_no_reingreso, var};
            datos_re = datos_tabla{grupo_reingreso, var};
            data = [datos_no; datos_re];
            group = [ones(length(datos_no).1);
2*ones(length(datos_re).1)];

            boxplot(data, group, 'Notch', 'off', 'Labels', {'No
Reingreso', 'Reingreso'});
            title(var, 'FontSize', 8, 'Interpreter', 'none');
            ylabel(var, 'FontSize', 8, 'Interpreter', 'none');
```

```
        end
    end
end
```

Cohorte prospectiva

```
% Cargo la base de datos:
datos_tabla = readtable('DatosRecogidos_FILTRADOS.xlsx',
'VariableNamingRule', 'preserve');

datos_tabla_origen = readtable('Reingresos_DatosRecogidos.xlsx');
variable_reingreso = {'Reingreso'}
datos_reingreso = datos_tabla_origen{:, variable_reingreso}

% Defino las variables de interés:
% Variables continuas:
variables_continuas = {'N_ingresos'};

% Variables categóricas:
variables_categoricas = {'I_Muco', 'A_Muco', ...
    'Microorg_Resis','Test_TAI_I_Inc',...
    'B_Oxi_Dom', 'A_Teo', 'Uso_Muscul_Acce', ...
    'Causa_Bacteriana', 'B_Teo', 'Gold', ...
    'I_Min_LABA', 'I_Min_SABA', 'Test_E5D_Cpe',...
    'Mov_Toracicos','Comp_Arritmias', ...
    'I_Teo', 'Anemia', 'A_Min_SABA', ...
    'Edemas_Per','I_VNI','B_Muco',...
    'Test_TAI_I_Err','Comp_Neumonia'};


datos_tabla.Reingreso = datos_reingreso;

grupo_no_reingreso = datos_tabla.Reingreso == 0; % Almaceno el grupo de
no reingreso.
grupo_reingreso = datos_tabla.Reingreso == 1; % Almaceno el grupo
reingreso.

% Calculo el número de pacientes de cada clase:
fprintf('Tamaño grupo "no reingreso": %d\n', sum(grupo_no_reingreso));
fprintf('Tamaño grupo "reingreso": %d\n', sum(grupo_reingreso));

% Inicializo la tabla de resultados:
resultados = table();

% Calculo los estadísticos descriptivos para las variables previas:
for i = 1:length(variables_continuas)
    variable_actual = variables_continuas{i};
```

```
    datos_no_reingreso = datos_tabla{grupo_no_reingreso,
variable_actual};
    datos_reingreso = datos_tabla{grupo_reingreso, variable_actual};

    datos_total = datos_tabla{:, variable_actual};

    % Me quedo con los que no contegan NaN:
    valid_no_reingreso = sum(~isnan(datos_no_reingreso));
    valid_reingreso = sum(~isnan(datos_reingreso));
    valid_total = sum(~isnan(datos_total));

    fprintf('Variable: %s, Datos válidos "no reingreso" : %d,
"reingreso": %d, Total: %d\n', ...
        variable_actual, valid_no_reingreso, valid_reingreso,
valid_total);

    % Calculo la mediana, primer y tercer cuartil:
    mediana_no_reingreso = NaN;
    q1_no_reingreso = NaN;
    q3_no_reingreso = NaN;

    mediana_reingreso = NaN;
    q1_reingreso = NaN;
    q3_reingreso = NaN;

    mediana_total = NaN;
    q1_total = NaN;
    q3_total = NaN;

    p_valor = NaN;

    if valid_no_reingreso > 0
        mediana_no_reingreso = median(datos_no_reingreso, 'omitnan');

        q1_no_reingreso = quantile(datos_no_reingreso, 0.25); % Primer
cuartil.
        q3_no_reingreso = quantile(datos_no_reingreso, 0.75); % Tercer
cuartil.
    end

    if valid_reingreso > 0
        mediana_reingreso = median(datos_reingreso, 'omitnan');

        q1_reingreso = quantile(datos_reingreso, 0.25); % Primer cuartil.
        q3_reingreso = quantile(datos_reingreso, 0.75); % Tercer cuartil.
    end

    if valid_total > 0
        mediana_total = median(datos_total, 'omitnan');
```

```
        q1_total = quantile(datos_total, 0.25); % Primer cuartil.
        q3_total = quantile(datos_total, 0.75); % Tercer cuartil.
    end

    if valid_no_reingreso >= 1 && valid_reingreso >= 1
        try
            [p_valor, ~]= ranksum(datos_no_reingreso, datos_reingreso); %
Test Mann-Whitney
        catch
            fprintf('Imposible aplicar Mann-Whitney para %s: faltan
datos\n', variable_actual);
        end
    end

    % Genero la tabla de estadísticos descriptivos:
    resultados =[resultados; table({variable_actual}, ...
        {[num2str(mediana_total, '%.1f') ' (' num2str(q1_total, '%.1f')
', ' num2str(q3_total, '%.1f') ')']}, ...
        {[num2str(mediana_no_reingreso, '%.1f') ' ('
num2str(q1_no_reingreso, '%.1f') ', ' num2str(q3_no_reingreso, '%.1f')
')']}, ...
        {[num2str(mediana_reingreso, '%.1f') ' (' num2str(q1_reingreso,
'%.1f') ', ' num2str(q3_reingreso, '%.1f') ')']}, ...
        {p_valor}, 'VariableNames', {'Variable', 'Total', 'No reingreso',
'Reingreso', 'p_valor'})];
end


% Calculo en forma de porcentaje los pacientes que hay en cada categoría:
for i = 1:length(variables_categoricas)
    variable = variables_categoricas{i};
    if ~iscategorical(datos_tabla.(variable))
        datos_tabla.(variable) = categorical(datos_tabla.(variable));
    end

    % Por cada categoría, calculo ese porcentaje:
    categorias = categories(datos_tabla.(variable));

    for j = 1:length(categorias)

        categoria = categorias{j};
        fprintf(variable + " "+ categoria +"\n");

        if ~(strcmp(variable, 'Test_E5D_Dol') && strcmp(categoria, '0'))

            sum_no_reingreso = sum(datos_tabla{grupo_no_reingreso,
variable} == categoria);
```

```
            sum_reingreso = sum(datos_tabla{grupo_reingreso, variable} ==
categoria);

            total_no_reingreso = sum(grupo_no_reingreso);
            total_reingreso = sum(grupo_reingreso);
            total_sample =height(datos_tabla);
            sum_total = sum(datos_tabla{:, variable} == categoria);

            porcentaje_no_reingreso = (sum_no_reingreso /
total_no_reingreso) * 100;
            porcentaje_reingreso = (sum_reingreso / total_reingreso) *
100;
            porcentaje_total = (sum_total / total_sample) * 100;

            tabla_contingencia =[sum_no_reingreso, total_no_reingreso -
sum_no_reingreso; ...
                                 sum_reingreso, total_reingreso -
sum_reingreso];

            p_valor = NaN;

            if any(tabla_contingencia(:) > 0) % Manejo de posibles datos
faltantes.
                try
                    [~, p_valor] = fishertest(tabla_contingencia); %
Aplico el Test de Fisher.
                catch
                    fprintf('Imposible aplicar el Test de Fisher para %s
(%s): faltan datos\n', variable, categoria);
                end
            end

            % Muestro los resultados del test en la tabla:
            resultados =[resultados; table({[variable ' (' categoria
')']}, ...
                {[num2str(sum_total) ' (' num2str(porcentaje_total,
'%.1f') '%)']}, ...
                {[num2str(sum_no_reingreso) ' ('
num2str(porcentaje_no_reingreso, '%.1f') '%)']}, ...
                {[num2str(sum_reingreso) ' ('
num2str(porcentaje_reingreso, '%.1f') '%)']}, ...
                {p_valor}, ...
                'VariableNames', {'Variable', 'Total', 'No reingreso',
'Reingreso', 'p_valor'})];

        end
    end
end
```

```
% Muestro los resultados de los estadísticos, tanto de variables
continuas como categóricas, en tabla:
disp(resultados);

% Guardo los resultados en un nuevo Excel:
writetable(resultados, 'Resultados_EstDescriptivos_prospectiva.xlsx');


% Filtrar variables continuas con p-valor < 0.05
significativas = {};
for i = 1:height(resultados)
    var = resultados.Variable{i};
    if ismember(var, variables_continuas) &&
~isnan(resultados.p_valor{i}) && resultados.p_valor{i} < 0.05
        significativas = [significativas; var];
    end
end

num_sig = length(significativas);

if num_sig == 0
    disp('No hay variables continuas con diferencias significativas.');
else
    % Calcular cuántas figuras se necesitan (máximo 2 subplots por
figura)
    num_figs = ceil(num_sig);
    for f = 1:num_figs
        figure;
        for s = 1
            idx = (f-1)+ s;
            if idx > num_sig
                break;
            end
            var = significativas{idx};

            % Preparar datos para el boxplot
            datos_no = datos_tabla{grupo_no_reingreso, var};
            datos_re = datos_tabla{grupo_reingreso, var};
            data = [datos_no; datos_re];
            group = [ones(length(datos_no).1);
2*ones(length(datos_re).1)];

            % Crear subplot y boxplot
            boxplot(data, group, 'Notch', 'off', 'Labels', {'No
Reingreso', 'Reingreso'});
            title(['Diagrama de cajas para la variable ' var],
'FontSize', 12, 'Interpreter', 'none');
            ylabel(['Valores de la variable ' var], 'FontSize', 12,
'Interpreter', 'none');
```

```
            xlabel('Grupos bajo estudio', 'FontSize', 12);

            if ~exist('C:\plot', 'dir')
                mkdir('C:\plot');
            end
            print(['C:\\plot\\graf' num2str(f) '.png'], '-dpng');
        end
    end
end
```

ANEXO 3. Código selección de variables predictoras

En el siguiente código, se presenta la selección de características predictoras mediante *ReliefF*, calculándose los pesos o relevancia por cada variable y almacenando aquellas que superasen el umbral establecido. Se aplica para *K* un valor de 5.

```
% Cargo los datos:
tabla = readtable('Datos_imputacionkVecinosAplicado_FINAL.xlsx');

datos_tabla_origen = readtable('DatosEPOC.xlsx');
datos_tabla_origen = datos_tabla_origen(datos_tabla_origen.('A_Exitus')
~= 1, :);
variable_reingreso = {'Reingreso'}
datos_reingreso = datos_tabla_origen{:, variable_reingreso}

tabla.Reingreso = datos_reingreso;

% Las variables predictoras se almacenan en X y la variable target se
guarda en Y:
Y_tabla = tabla(:, 'Reingreso'); %Variable target

nom_variables = tabla.Properties.VariableNames;
indice_col_reingreso = find(strcmp(nom_variables, 'Reingreso'));
indices_col = 1:width(tabla);
incides_var_predictoras = setdiff(indices_col, indice_col_reingreso);

% Variables predictoras
X_tabla = tabla(:, incides_var_predictoras);


X = table2array(X_tabla);
Y = table2array(Y_tabla);

X_std = zscore(X); % Estandarizo las variables predictoras almacenadas en
X.
k = 5; %Número de vecinos
[indices, weights] = relieff(X_std, Y, k); % Proporciona los índices
ordenados y los pesos/puntuaciones de importancia sobre el conjunto.


% Guardo los nombres de las variables predictoras (X) y de esos nombres
ordenados según la puntuación obtenida (relevancia).

nombres_variables = X_tabla.Properties.VariableNames; % Sin variable
Reingreso.
nombres_ordenados = nombres_variables(indices); % Orden atendiendo a su
relevancia.
```

```
pesos_ordenados = weights(indices); % Puntuaciones ordenadas de mayor a
menor.
pesos_positivos = [] % Inicializo.
for i = 1:length(pesos_ordenados)
    if pesos_ordenados(i)> 0
        pesos_positivos(i) = pesos_ordenados(i) % Guardo solo los pesos
positivos.
    end
end

media_pesos = mean(pesos_positivos) % Calculo la media de esos pesos
positivos obtenidos para facilitar la elección del valor de "K" tras
varias
% pruebas, pudiendo ver cuál de las "K" devuelve un promedio con valores
más cercanos a 1.

% Me quedo con las 24 variables de mayor peso, cumpliendo así la regla
empírica de 10 instancias por variable.
num_variables = 24;

% Genero una tabla con los nombres de las variables y sus pesos ordenados
de mayor a menor.
tabla_importancia_5 = table(nombres_ordenados', pesos_ordenados', ...
    'VariableNames', {'Variable', 'Peso'});
disp('Tabla de variables ordenadas por puntuación (peso):');
disp(tabla_importancia_5);

tabla_seleccionada_5 = tabla_importancia_5(1:num_variables, :); % Me
quedo únicamente con las 24 variables de mayor peso.
disp('Primeras 24 variables predictoras seleccionadas:');
disp(tabla_seleccionada_5);

% Genero una matriz que contenga únicamente esas variables.
indices_seleccionados = indices(1:num_variables);
X_selected = X(:, indices_seleccionados);
nombres_seleccionados = nombres_ordenados(1:num_variables);
tabla_X_selected = array2table(X_selected, 'VariableNames',
nombres_seleccionados);

% Guardo los resultados en Excel.
writetable(tabla_X_selected, 'variablesElegidasRelief_K5.xlsx');

% Creo un gráfico de barras que muestre las 24 variables seleccionadas y
sus pesos.
figure;
bar(tabla_seleccionada_5.Peso);
nombres_con_espacios = strrep(tabla_seleccionada_5.Variable, '_', ' ');

set(gca, 'XTick', 1:num_variables, 'XTickLabel', nombres_con_espacios);
```

```
xtickangle(45);
xlabel('Variables');
ylabel('Puntuación (Peso)');
title('Primeras 24 variables predictoras seleccionadas con K = 5');
grid on;
```

ANEXO 4. Código modelo predictivo basado en *Random Forest.*

Se adjunta el *script* que contiene el modelo predictivo basado en *Random Forest.* En primer lugar, se almacena la variable *target* y aquellas seleccionadas finalmente por *ReliefF*, que resultaron ser las 24 variables con los mayores pesos. Tras esto, puede observarse la partición de los datos en *test y train1.* Este último se divide a su vez en *train2* y *val*. Estos cuatro conjuntos se estandarizan y se realiza la optimización de los hiperparámetros y umbral de decisión. Para obtener el mejor valor para dichos hiperparámetros, se establece como criterio de selección la maximización de *F1 score*. Después, se reentrena el modelo final con todos los hiperparámetros ya optimizados mediante el conjunto *train1* y finalmente se aplican diversas métricas de rendimiento.

```
% En primer lugar, se obtiene la variable target (objetivo) de la tabla:
tabla_partida = readtable('Datos_imputacionkVecinosAplicado_FINAL.xlsx');

datos_tabla_origen = readtable('DatosEPOC.xlsx');
datos_tabla_origen = datos_tabla_origen(datos_tabla_origen.('A_Exitus')
~= 1, :);
variable_reingreso = {'Reingreso'}
datos_reingreso = datos_tabla_origen{:, variable_reingreso}
Y_target = datos_reingreso;

% Cargo los datos con las variables seleccionadas:
tabla_variables_seleccionadas =
readtable("variablesElegidasRelief_K5.xlsx");
tabla_variables_seleccionadas = tabla_variables_seleccionadas(:, 1:24); %
Selecciono las 24 primeras variables predictoras.
matriz_variables_seleccionadas =
table2array(tabla_variables_seleccionadas); % Convierto la tabla en
matriz.


% PRIMERA DIVISIÓN DE LOS DATOS EN train1 y test:
cv_1 = cvpartition(Y_target, 'HoldOut', 0.3); % El 70% de los datos:
entrenamiento. El resto (30%): test.
idx_Train1 = training(cv_1); % Índices de los datos que pertenecen a
entrenamiento.
idx_Test = test(cv_1); % Índices de los datos que pertenecen a test.

X_train1 = matriz_variables_seleccionadas(idx_Train1,:);
Y_train1 = Y_target(idx_Train1);

X_test = matriz_variables_seleccionadas(idx_Test,:);
Y_test = Y_target(idx_Test);

% SEGUNDA PARTICIÓN SOBRE train1:
```

```
cv_2 = cvpartition(Y_train1, "HoldOut", 0.3); % Vuelvo a dividir los
datos de la misma manera: 70% train2 y 30% para el conjunto val.
idx_Train2 = training(cv_2); % Índices de los datos que pertenecen a
entrenamiento.
idx_Val = test(cv_2); % Índices de los datos que pertenecen a test.

X_train2 = X_train1(idx_Train2, :);
Y_train2 = Y_train1(idx_Train2);

X_valid = X_train1(idx_Val, :);
Y_val = Y_train1(idx_Val);

% Estandarizo:
mu = mean(X_train2);
sigma = std(X_train2);
% Para evitar que se produzacn divisiones entre 0, se reemplazan las
desviaciones nulas por 1.
sigma(sigma == 0) = 1;
X_train2 = (X_train2 - mu) ./ sigma;
X_valid = (X_valid - mu) ./ sigma;
X_test = (X_test - mu) ./ sigma;
X_train1 = (X_train1 - mu) ./ sigma;

%-------------------------------------------------------------------------
----------------
% OPTIMIZACIÓN DE LOS HIPERPARÁMETROS:
% Loas hiperparámetrros a optimizar son: meanleafsize, numpredictors,
maxnumsplits, num_arboles y penalizaciones de falsos positivos y falsos
negativos.

%1º OPTIMIZACIÓN: numero de arboles
% Valores fijos supuestos:
NumPredictorsToSample = 5; % Número de predictores a muestrear.
MaxNumSplits = 20; % Máximo número de divisiones.
MinLeafSize = 5; % Tamaño mínimo de hoja.
matrizCostes = [0 5; 10 0]; % [coste TN, coste FP; coste FN, coste TP].
Se le otorga mayor penalización a los falsos negativos, para mitigar
%el desbalance de clases.

% Rango de posibles valores para el número de árboles:
numTrees = 5:10:400;

F1_scores_trees = zeros(size(numTrees)); % Inicializo F1 score.

% Pruebo los distintos números de árboles, seleccionando el que maximice
F1 score:
for i = 1:length(numTrees)
    nTrees = numTrees(i);
    RandomF = TreeBagger(nTrees, X_train2, Y_train2, ...
```

```
                        'Method', 'classification', ...
                        'NumPredictorsToSample', NumPredictorsToSample,
...
                        'MaxNumSplits', MaxNumSplits, ...
                        'MinLeafSize', MinLeafSize, ...
                        'Cost', matrizCostes);

    [pred_trees, scores] = RandomF.predict(X_valid);
    pred_trees = str2double(pred_trees);

    % Me aseguro de que las predicciones se hayan realizado
correctamente:
    if any(isnan(pred_trees)) || isempty(pred_trees)
        warning('Predicciones no válidas para el número de árboles = %d',
nTrees);
        F1_scores_trees(i) = 0;
        continue;
    end

    % A continuación, calculo F1 score. Para ello, necesito primero
obtener la matriz de confusión:
    matrizConfusion_trees = confusionmat(Y_val, pred_trees,'Order', [0
1]); % Matriz de confusión
    if sum(sum(matrizConfusion_trees)) == 0 %Me aseguro de que la matriz
recién creada no presente valores vacíos.
        warning('Matriz de confusión sin valores con un número de árboles
de = %d', nTrees);
        F1_scores_trees(i) = 0;
        continue;
    end

    fprintf('numTrees=%d', nTrees);
    F1_scores_trees(i) = Calc_f1Score(matrizConfusion_trees);
end

[bestNumTrees, maxF1_trees] = Plot_f1_param(numTrees, F1_scores_trees,
...
    'Número de árboles', 'Optimización del hiperparámetro numTrees',
'Número de árboles');


%-----------------------------------------------------------------------
----------------
%2º OPTIMIZACIÓN: penalización para falsos positivos.
costFP_values = 2:0.5:15;  % Pruebo valores de de 2 a 15 en pasos de 0.5.
F1_score_FPcost = zeros(size(costFP_values));

% Entreno el modelo para cada valor de prueba, integrando ya el número de
árboles optimizado:
```

```
for i = 1:length(costFP_values)
    cost_FP = costFP_values(i);
    matriz_costes_op_FP = [0 cost_FP; 10 0];  % Fijo una penalización de
10 para falsos negativos.


    RF = TreeBagger(bestNumTrees, X_train2, Y_train2, ...
                    'Method', 'classification', ...
                    'NumPredictorsToSample', NumPredictorsToSample, ...
                    'MaxNumSplits', MaxNumSplits, ...
                    'MinLeafSize', MinLeafSize, ...
                    'Cost', matriz_costes_op_FP);

    % Realizo la predicción sobre el conjunto de validación:
    [predCost_FP, ~] = RF.predict(X_valid);
    predCost_FP = str2double(predCost_FP);

    % Me aseguro de que las predicciones se hayan realizado
correctamente:
    if any(isnan(predCost_FP)) || isempty(predCost_FP)
        warning('Predicciones no válidas para una penalización de FP de =
%.2f', cost_FP);
        F1_score_FPcost(i) = 0;
        continue;
    end

    % Calculo F1 score, pero primero necesito la matriz de confusión:
    matrizConfusionCost_FP = confusionmat(Y_val, predCost_FP, 'Order', [0
1]);
    if sum(sum(matrizConfusionCost_FP)) == 0 % Me aseguro de que la
matriz recién creada no presente valores vacíos.
        warning('Matriz de confusión con valores faltantes para una
penalización de FP = %.2f', cost_FP);
        F1_score_FPcost(i) = 0;
        continue;
    end

    fprintf('costFP=%.2f', cost_FP);
    F1_score_FPcost(i) = Calc_f1Score(matrizConfusionCost_FP);
end

[bestCostFP, maxF1_FP] = Plot_f1_param(costFP_values, F1_score_FPcost,
...
    'Coste de Falso Positivo', 'Optimización del hiperparámetro: Coste
FP', 'Penalización de FP');


%-------------------------------------------------------------------------
----------------
```

```
%3º OPTIMIZACIÓN: penalización para falsos negativos.
costFN_values = 16:0.5:30;  % Pruebo valores de 16 a 30 en pasos de 0.5.
F1_score_FNcost = zeros(size(costFN_values));

% Entreno de nuevo, pero esta vez con el número de árboles y penalización
de FP ya optimizados:
for i = 1:length(costFN_values)
    cost_FN = costFN_values(i);
    matriz_costes_op_FN = [0 bestCostFP; cost_FN 0];


    RF = TreeBagger(bestNumTrees, X_train2, Y_train2, ...
                    'Method', 'classification', ...
                    'NumPredictorsToSample', NumPredictorsToSample, ...
                    'MaxNumSplits', MaxNumSplits, ...
                    'MinLeafSize', MinLeafSize, ...
                    'Cost', matriz_costes_op_FN);


    [predCost_FN, ~] = RF.predict(X_valid);
    predCost_FN = str2double(predCost_FN);


    if any(isnan(predCost_FN)) || isempty(predCost_FN)
        warning('Predicciones no válidas para una penalización de FN de =
%.2f', cost_FN);
        F1_score_FNcost(i) = 0;
        continue;
    end

    % Calculo la matriz de confusión para luego obtener F1 score:
    matrizConfusionCost_FN = confusionmat(Y_val, predCost_FN, 'Order', [0
1]);
    if sum(sum(matrizConfusionCost_FN)) == 0
        warning('Matriz de confusión con datos vacíos para una
penalización de FN = %.2f', cost_FN);
        F1_score_FNcost(i) = 0;
        continue;
    end

    fprintf('costFN=%.2f', cost_FN);
    F1_score_FNcost(i) = Calc_f1Score(matrizConfusionCost_FN);

end

[bestCostFN, maxF1_FN] = Plot_f1_param(costFN_values, F1_score_FNcost,
...
    'Coste de Falso Negativo', 'Optimización del hiperparámetro: Coste
FN', 'Penalización de FN');
```

```
%-------------------------------------------------------------------------
---------------
% 4º OPTIMIZACIÓN: MinLeafSize (tamaño mínimo de hoja)
NumPredictorsToSample = 5;
MaxNumSplits = 20;
matrizCostes_op = [0 bestCostFP; bestCostFN 0]; % Integro para el
entrenamiento las penalizaciones ya optimizadas.

minLeafSize_valores = 5:1:15; % Pruebo valores de 5 a 15 de 1 en 1.
F1_scores_leaf = zeros(size(minLeafSize_valores));

% Entreno con las penalizaciones ya optimizadas y el número de árboles:
for i = 1:length(minLeafSize_valores)
    min_hoja = minLeafSize_valores(i);

    RF = TreeBagger(bestNumTrees, X_train2, Y_train2, ...
                    'Method', 'classification', ...
                    'NumPredictorsToSample', NumPredictorsToSample, ...
                    'MaxNumSplits', MaxNumSplits, ...
                    'MinLeafSize', min_hoja, ...
                    'Cost', matrizCostes_op);


    [pred_leaf, ~] = RF.predict(X_valid);
    pred_leaf = str2double(pred_leaf);


    if any(isnan(pred_leaf)) || isempty(pred_leaf)
        warning('Predicciones no válidas para un tamaño mínimo de hoja de
= %d', min_hoja);
        F1_scores_leaf(i) = 0;
        continue;
    end

    % Calculo la matriz de confusión:
    matrizConfusion_leaf = confusionmat(Y_val, pred_leaf, 'Order', [0
1]);
    if sum(sum(matrizConfusion_leaf)) == 0
        warning('Matriz de confusión vacía para MinLeafSize = %d',
min_hoja);
        F1_scores_leaf(i) = 0;
        continue;
    end

    fprintf('MinLeafSize=%d', min_hoja);
    F1_scores_leaf(i) = Calc_f1Score(matrizConfusion_leaf);
end
```

```
[bestMinLeafSize, maxF1_leaf] = Plot_f1_param(minLeafSize_valores,
F1_scores_leaf, ...
    'MinLeafSize', 'Optimización del hiperparámetro: MinLeafSize',
'Tamaño mínimo de hoja');


%-------------------------------------------------------------------
----------------------------------
% 5º OPTIMIZACIÓN: número de predictores
MaxNumSplits = 20;
num_features = size(X_train2, 2);
predictor_values = 2:1:num_features;
F1_predictor = zeros(size(predictor_values));

for i = 1:length(predictor_values)
    nPred = predictor_values(i);

    RF = TreeBagger(bestNumTrees, X_train2, Y_train2, ...
                    'Method', 'classification', ...
                    'NumPredictorsToSample', nPred, ...
                    'MinLeafSize', bestMinLeafSize, ...
                    'MaxNumSplits', MaxNumSplits, ...
                    'Cost', [0 bestCostFP; bestCostFN 0]);

    [pred_predictors, ~] = RF.predict(X_valid);
    pred_predictors = str2double(pred_predictors);


    if any(isnan(pred_predictors)) || isempty(pred_predictors)
        warning('Predicciones no válidas para un número de predictores de
= %d', nPred);
        F1_predictor(i) = 0;
        continue;
    end

    % Matriz de confusión:
    confusion_pred = confusionmat(Y_val, pred_predictors, 'Order', [0
1]);
    if sum(sum(confusion_pred)) == 0
        warning('Matriz de confusión con datos faltantes para un núemeero
de predictores = %d', nPred);
        F1_predictor(i) = 0;
        continue;
    end

    fprintf('NumPredictors=%d', nPred);
    F1_predictor(i) = Calc_f1Score(confusion_pred);
end
```

```
[bestNumPredictors, maxF1_pred] = Plot_f1_param(predictor_values,
F1_predictor, ...
    'NumPredictorsToSample', 'Optimización de NumPredictorsToSample',
'Número de predictores');


%-------------------------------------------------------------------
--
% 6º OPTIMIZACIÓN: número máximo de divisiones (MaxNumSplits)
maxSplits_range = 10:10:100; % Pruebo valores en este rango.
F1_maxSplits = zeros(size(maxSplits_range));

for i = 1:length(maxSplits_range)
    maxSplits = maxSplits_range(i);

    RF = TreeBagger(bestNumTrees, X_train2, Y_train2, ...
                    'Method', 'classification', ...
                    'NumPredictorsToSample', bestNumPredictors, ...
                    'MinLeafSize', bestMinLeafSize, ...
                    'MaxNumSplits', maxSplits, ...
                    'Cost', [0 bestCostFP; bestCostFN 0]);

    [pred_splits, ~] = RF.predict(X_valid);
    pred_splits = str2double(pred_splits);


    if any(isnan(pred_splits)) || isempty(pred_splits)
        warning('Predicciones no válidas para un número máximo de
divisiones = %d', maxSplits);
        F1_maxSplits(i) = 0;
        continue;
    end

    % Matriz de confusión:
    matrizConfusion_splits = confusionmat(Y_val, pred_splits, 'Order', [0
1]);
    if sum(sum(matrizConfusion_splits)) == 0
        warning('Matriz de confusión con datos vacíos para un número
máximo de divisiones = %d', maxSplits);
        F1_maxSplits(i) = 0;
        continue;
    end

    fprintf('MaxNumSplits=%d', maxSplits);
    F1_maxSplits(i) = Calc_f1Score(matrizConfusion_splits);
end

[bestMaxSplits, maxF1_splits] = Plot_f1_param(maxSplits_range,
F1_maxSplits, ...
```

```
    'MaxNumSplits', 'Optimización del hiperparámetro: MaxNumSplits',
'Número de divisiones');


%-------------------------------------------------------------------------
-----------------------------
% 7º OPTIMIZACIÓN: Umbral de decisión
% Una vez obtenidos los hiperparámetros óptimos del árbol, se optimiza el
umbral de predicción:
modeloVal = TreeBagger(bestNumTrees, X_train1, Y_train1, ...
                       'Method', 'classification', ...
                       'NumPredictorsToSample', bestNumPredictors, ...
                       'MinLeafSize', bestMinLeafSize, ...
                       'MaxNumSplits', bestMaxSplits, ...
                       'Cost', [0 bestCostFP; bestCostFN 0]);


[~, scoresVal] = modeloVal.predict(X_valid);
scoresVal = scoresVal(:, 2); % Probabilidad de obtener la clase positiva.

% Obtengo la curva ROC:
[X_roc, Y_roc, T_roc, AUC] = perfcurve(Y_val, scoresVal, 1);

% Obtengo F1 Score para cada umbral mediante la curva ROC recién
obtenida.
F1_scores_umbral = zeros(size(T_roc));
for i = 1:length(T_roc)
    thresh = T_roc(i);
    pred_umbral = double(scoresVal >= thresh);


    if any(isnan(pred_umbral)) || isempty(pred_umbral)
        warning('Predicciones no válidas para un umbral de = %.2f',
thresh);
        F1_scores_umbral(i) = 0;
        continue;
    end

    % Matriz de confusión:
    matrizConfusion_umbral = confusionmat(Y_val, pred_umbral, 'Order', [0
1]);
    if sum(sum(matrizConfusion_umbral)) == 0
        warning('Matriz de confusión con daltos faltantes para un umbral
de = %.2f', thresh);
        F1_scores_umbral(i) = 0;
        continue;
    end

    fprintf('umbral=%.2f', thresh);
```

```
    F1_scores_umbral(i) = Calc_f1Score(matrizConfusion_umbral);
end


% Elimino los NaNs:
valid_idx = ~isnan(F1_scores_umbral);
T_roc_clean = T_roc(valid_idx);
F1_scores_umbral_clean = F1_scores_umbral(valid_idx);
X_roc_clean = X_roc(valid_idx);
Y_roc_clean = Y_roc(valid_idx);

% Me quedo con el umbral que maximiza el F1 score:
[maxF1_umbral, idxMaxF1_umbral] = max(F1_scores_umbral_clean);
mejor_umbral = T_roc_clean(idxMaxF1_umbral);
bestX = X_roc_clean(idxMaxF1_umbral);
bestY = Y_roc_clean(idxMaxF1_umbral);

% Represento la curva ROC:
figure;
plot(X_roc_clean, Y_roc_clean, 'b-', 'LineWidth', 2);
hold on;

% Marco en la curva ROC el mejor punto de F1 score:
plot(bestX, bestY, 'ro', 'MarkerSize', 8, 'LineWidth', 2);

xlabel('Tasa de falsos positivos (FPR)');
ylabel('Tasa de verdaderos positivos (TPR)');
title('Curva ROC y umbral óptimo');
legend('Curva ROC', 'Máximo F1', 'Location', 'southeast');
grid on;
hold off;

fprintf('Umbral de decisión optimizado = %.2f con F1 score = %.4f\n',
mejor_umbral, maxF1_umbral);
fprintf('AUC obtenido por el modelo: %.4f\n', AUC);

%-------------------------------------------------------------------------
-----
% Evaluación del modelo ya optimizado y con el úmbral también optimizado:
modelo_def = TreeBagger(bestNumTrees, X_train1, Y_train1, ...
                        'Method', 'classification', ...
                        'NumPredictorsToSample', bestNumPredictors, ...
                        'MinLeafSize', bestMinLeafSize, ...
                        'MaxNumSplits', bestMaxSplits, ...
                        'Cost', [0 bestCostFP; bestCostFN 0]);

% Ahora las predicciones se realizan sobre el conjunto test:
[~, scoresDef] = modelo_def.predict(X_test);
scoresDef = scoresDef(:, 2); % Probabilidad de obtener la clase positiva.
```

```
% Obtener la curva ROC y AUC:
[fpr, tpr, ~, AUC] = perfcurve(Y_test, scoresDef, 1);
fprintf('AUC del modelo final: %.4f\n', AUC);

% Curva ROC:
figure;
plot(fpr, tpr, 'LineWidth', 2);
xlabel('1 - Especificidad (FPR)');
ylabel('Sensibilidad (TPR)');
title(sprintf('Curva ROC (AUC = %.4f)', AUC));
grid on;
predTest = double(scoresDef >= mejor_umbral);


if any(isnan(predTest)) || isempty(predTest)
    error('Predicciones no válidas en el conjunto test');
end

% Matriz de confusión:
matrizConfusion_final = confusionmat(Y_test, predTest, 'Order', [0 1]);
if sum(sum(matrizConfusion_final)) == 0
    error('Matriz de confusión con datos faltantes en el conjunto de
test');
end

TN = matrizConfusion_final(1.1);
FP = matrizConfusion_final(1.2);
FN = matrizConfusion_final(2.1);
TP = matrizConfusion_final(2.2);


se = TP / (TP + FN);
sp = TN / (TN + FP);
acc = (TP + TN) / (TP + FN + TN + FP); % accuracy
ppv = TP / (TP + FP);
F1_score_final = 2 * ((ppv * se) / (ppv + se));
LR_pos = se / (1 - sp); % Relación de verosimilitud positiva.
LR_neg = (1 - se) / sp; % Relación de verosimilitud negativa.
npv = TN / (TN + FN); % Valor predictivo negativo.

fprintf('\n--- RESULTADOS SOBRE TEST SET (UMBRAL OPTIMIZADO) ---\n');
fprintf('Sensibilidad: %.4f\n', se);
fprintf('Especificidad: %.4f\n', sp);
fprintf('Precisión (accuracy): %.4f\n', acc);
fprintf('F1 score: %.4f\n', F1_score_final);
fprintf('LR positiva: %.4f\n', LR_pos);
fprintf('LR negativa: %.4f\n', LR_neg);
```

```
fprintf('NPV: %.4f\n', npv);
fprintf('PPV: %.4f\n', ppv);
disp('Matriz de confusión:');
disp(matrizConfusion_final);


function [F1_score] = Calc_f1Score(matrizConfusion_umbral)
    % Función que calcula F1 score a partir de la
    % matriz de confusion
    TN = matrizConfusion_umbral(1.1);
    FP = matrizConfusion_umbral(1.2);
    FN = matrizConfusion_umbral(2.1);
    TP = matrizConfusion_umbral(2.2);

    se = TP / (TP + FN);
    ppv = TP / (TP + FP);
    F1_score = 2 * ((se * ppv) / (se + ppv));

    fprintf('TP=%d, FP=%d, FN=%d, TN=%d, se=%.4f, ppv=%.4f, F1=%.4f\n',
...
            TP, FP, FN, TN, se, ppv, F1_score);
end

function [bestParam, maxF1] = Plot_f1_param(valores_x, valores_f1,
xLabelStr, titulo_plot, nombre_parametro)
    % Función para representar la evolución de f1 score
    % frente al rango de hiperpárametros a provar
    valid_idx = ~isnan(valores_f1);
    x_clean = valores_x(valid_idx);
    F1_clean = valores_f1(valid_idx);

    figure;
    plot(x_clean, F1_clean, '-o', 'LineWidth', 1.5);
    hold on;

    % Busca el valor óptimo
    [maxF1, idxMax] = max(F1_clean);
    bestParam = x_clean(idxMax);

    % Marca el punto óptimo
    plot(bestParam, maxF1, 'ro', 'MarkerSize', 8, 'LineWidth', 2);

    % Etiquetas
    xlabel(xLabelStr);
    ylabel('F1 Score');
    title(titulo_plot);
    legend('F1 Score', 'Máximo F1', 'Location', 'best');
    grid on;
    hold off;
```

```
    % Imprime resultados
    if isnumeric(bestParam) && mod(bestParam,1)==0
        fprintf('%s óptimo = %d con F1 score = %.4f\n', nombre_parametro,
bestParam, maxF1);
    else
        fprintf('%s óptimo = %.2f con F1 score = %.4f\n',
nombre_parametro, bestParam, maxF1);
    end
end
```

ANEXO 5. Código modelo predictivo basado en red neuronal perceptrón multicapa.

Análogamente al anterior *script,* se realizan las mismas particiones de datos expuestas para *Random Forest*. Se estandarizan los mismos y se lleva a cabo la optimización de dos hiperparámetros de la red: el número de neuronas de la capa oculta y el parámetro de regularización (*alpha*). Se genera una matriz cuyos elementos son el *F1 score* calculado para cada combinación de los hiperparámetros. De nuevo, la combinación óptima será aquella que maximice *F1 score.* También se adjunta la optimización del umbral de decisión. Por último, se reentrena el modelo con *train1* aplicando todos los parámetros ya optimizados y finalmente se calculan las métricas de rendimiento para evaluar la red neuronal.

```
% Cargo la base de datos y extraigo de ella la variable reingreso
(target):
tabla_partida = readtable('Datos_imputacionkVecinosAplicado_FINAL.xlsx');

datos_tabla_origen = readtable('DatosEPOC.xlsx');
datos_tabla_origen = datos_tabla_origen(datos_tabla_origen.('A_Exitus')
~= 1, :);
variable_reingreso = {'Reingreso'}
datos_reingreso = datos_tabla_origen{:, variable_reingreso}
tabla_partida.Reingreso = datos_reingreso;

Y_target = tabla_partida.Reingreso;

tabla_variables_seleccionadas =
readtable("variablesElegidasRelief_K5.xlsx"); % Cargo el Excel que
contiene las variables que superaban el umbral.
tabla_variables_seleccionadas = tabla_variables_seleccionadas(:, 1:24); %
De las variables resultantes de ReliefF, me quedo con las 24 de mayor
peso.
matriz_variables_seleccionadas =
table2array(tabla_variables_seleccionadas);

% Creación de los conjuntos:
cv_1 = cvpartition(Y_target, 'HoldOut', 0.3); % Entrenamiento: 70%. Test:
30%.
idx_Train1 = training(cv_1);
idx_Test = test(cv_1);

X_train1 = matriz_variables_seleccionadas(idx_Train1,:);
Y_train1 = Y_target(idx_Train1);
X_test = matriz_variables_seleccionadas(idx_Test,:);
Y_test = Y_target(idx_Test);

cv_2 = cvpartition(Y_train1, 'HoldOut', 0.3);
```

```
idx_Train2 = training(cv_2);
idx_Val = test(cv_2);

X_train2 = X_train1(idx_Train2, :);
Y_train2 = Y_train1(idx_Train2);
X_val = X_train1(idx_Val, :);
Y_val = Y_train1(idx_Val);

% Estandarizo los conjuntos:
mu = mean(X_train2);
sigma = std(X_train2);
X_train2_scaled = (X_train2 - mu) ./ sigma;
X_val_scaled = (X_val - mu) ./ sigma;

% OPTIMIZACIÓN DE LOS HIPERPARÁMETOS: parámetro de regularización y
número de neuronas en la capa oculta.
regularizacion = 0.1:0.05:1;
num_neuronas = 2:1:50;

f1_matrix = NaN(length(num_neuronas), length(regularizacion));
F1_max = 0;

% Busco la combinación de valores de los dos parámetros que me maximice
F1 score:
for i = 1:length(num_neuronas)
    for j = 1:length(regularizacion)
        net = patternnet(num_neuronas(i));
        % Funciones de activación aplicadas en la capa oculta y la de
salida:
        net.layers{1}.transferFcn = 'tansig'; % Capa oculta
        net.layers{2}.transferFcn = 'logsig'; % Capa de salida
        net.performParam.regularization = regularizacion(j); % Parámetro
de regularización.
        net.trainParam.showWindow = false; % Instrucción para no mostrar
la ventana de entrenamiento.
        net = train(net, X_train2_scaled', Y_train2'); %Entrenamiento de
la red.

        Y_val_pred_prob = net(X_val_scaled');
        Y_val_pred = double(Y_val_pred_prob >= 0.5);

        C = confusionmat(Y_val, Y_val_pred); %Calculo la matriz de
confusión
        TN = C(1.1); % Verdaderos negativos.
        FP = C(1.2); % Falsos positivos.
        FN = C(2.1); % Falsos negativos.
        TP = C(2.2); % Verdaderos positivos.

        se = TP/(TP + FN); % Sensibilidad.
```

```
        ppv = TP/(TP + FP); % Valor Predictivo Positivo.
        F1 = 2 * (se * ppv)/(se + ppv);  % Calculo la F1 score.

        if ~isnan(F1) % Manejo los posibles NaNs.
            f1_matrix(i, j) = F1 * 100; %Alamceno los F1 score de cada
combinación en la matriz como porcentaje.

            if F1 > F1_max
                F1_max = F1;
                num_neuronas_opt = num_neuronas(i);
                reg_opt = regularizacion(j);
            end

        end

    end
end

% Creación de una gráfica que muestre el F1 score (eje Y), número de
neuoronas (eje X) y parámetro de regularización (alpha).
figure;
hold on;
colors = lines(length(regularizacion));

for j = 1:length(regularizacion)
    plot(num_neuronas, f1_matrix(:, j), '-^', 'LineWidth', 1.5, ...
        'Color', colors(j,:), ...
        'DisplayName', sprintf('alpha = %.3f', regularizacion(j)));
end

hold off;
xlabel('Número de nodos en la capa oculta');
ylabel('F1 Score (%)');
title('F1 Score para distintas combinaciones de hiperparámetros');
legend('Location', 'best');
grid on;

% Optimizo el umbral de decisión:
net_opt = patternnet(num_neuronas_opt);
net_opt.layers{1}.transferFcn = 'tansig'; % Funciones de activación.
net_opt.layers{2}.transferFcn = 'logsig';
net_opt.performParam.regularization = reg_opt;
net_opt.trainParam.showWindow = false;
net_opt= train(net_opt, X_train2_scaled', Y_train2');

Y_val_pred_prob = net_opt(X_val_scaled');
umbrales = 0.3:0.01:1; % Rango de umbrales.
F1_umbrales = NaN(size(umbrales)); %Inicializo F1 score.
```

```
for k = 1:length(umbrales)
    Y_val_pred = double(Y_val_pred_prob >= umbrales(k));
    C = confusionmat(Y_val, Y_val_pred);  % Matriz de confusión.
    TN = C(1.1); % Verdaderos negativos.
    FP = C(1.2); % Falsos positivos.
    FN = C(2.1); % Falsos negativos.
    TP = C(2.2); % Verdaderos positivos.

    se = TP/(TP + FN); % Sensibilidad.
    ppv = TP/(TP + FP); % Valor Predictivo Positivo.
    F1 = 2 * (se * ppv)/(se + ppv); % Calculo la F1 score.

    if ~isnan(F1)
        F1_umbrales(k) = F1;
    end

end

[maxF1_val, idx_umbral] = max(F1_umbrales, [], 'omitnan'); % Me quedo con
el umbral que maximina F1 score.
umbral_opt = umbrales(idx_umbral);

% Vuelvo a entrenar, pero con train1:
X_train1_scaled = (X_train1 - mu) ./ sigma;
X_test_scaled= (X_test - mu) ./ sigma;

net_final = patternnet(num_neuronas_opt);
net_final.layers{1}.transferFcn = 'tansig'; % Funciones de activación.
net_final.layers{2}.transferFcn = 'logsig';
net_final.performParam.regularization = reg_opt;
net_final.trainParam.showWindow = true;
net_final = train(net_final, X_train1_scaled', Y_train1');

% Predicción sobre el conjunto test:
Y_test_pred_prob = net_final(X_test_scaled');
Y_test_pred = double(Y_test_pred_prob >= umbral_opt);

C_test = confusionmat(Y_test, Y_test_pred);
TN = C_test(1.1);
FP = C_test(1.2);
FN = C_test(2.1);
TP = C_test(2.2);

% Métricas de rendimiento:
Se = TP/(TP + FN);
Sp = TN/(TN + FP);
PPV = TP/(TP + FP);
NPV = TN/(TN + FN); % Valor predictivo negativo.
LR_pos = Se/(1 - Sp); % Razón de verosimilitud positiva.
```

```
LR_neg =(1 - Se)/(Sp); % Razón de vero similitud negativa.
Accuracy = (TP + TN)/(TP + FN + TN + FP); % Precisión.
F1 = 2 * (Se * PPV)/(Se + PPV);

[X_roc, Y_roc, ~, AUC] = perfcurve(Y_test, Y_test_pred_prob, 1); %Calculo
la Curva ROC y AUC.

% Muestro los reusltados finales:
fprintf('-- Hiperparámetros óptimos --\n');
fprintf('Neuronas ocultas en la capa oculta: %d\n', num_neuronas_opt);
fprintf('Parámetro de regularización (alpha): %.3f\n', reg_opt);
fprintf('Umbral óptimo: %.2f\n', umbral_opt);
fprintf('F1 score sobre el conjunto de validación con umbral óptimo:
%.3f\n', maxF1_val);

fprintf('\n-- Resultados sobre el conjunto TEST --\n');
disp(C_test);
fprintf('Sensibilidad: %.3f\n', Se);
fprintf('Especificidad: %.3f\n', Sp);
fprintf('PPV: %.3f\n', PPV);
fprintf('NPV: %.3f\n', NPV);
fprintf('LR+: %.3f\n', LR_pos);
fprintf('LR-: %.3f\n', LR_neg);
fprintf('Accuracy: %.3f\n', Accuracy);
fprintf('F1 score: %.3f\n', F1);
fprintf('AUC (área bajo la curva): %.3f\n', AUC);

% Genero la Curva ROC:
figure;
plot(X_roc, Y_roc, 'b-', 'LineWidth', 2);
hold on;
plot([0 1], [0 1], 'k--');
xlabel('1 - Especificidad (FPR)');
ylabel('Sensibilidad (TPR)');
title(sprintf('Curva ROC (AUC = %.3f)', AUC));
grid on;
axis square;
```

ANEXO 6. Validación temporal prospectiva del modelo predictivo basado en *Random Forest*

A continuación, se muestra la validación temporal de *Random Forest* mediante la base de datos prospectiva recogida. Su implementación se efectúa sobre el modelo ya optimizado y se calculan diversas métricas de rendimiento con el fin de determinar la capacidad predictiva del modelo en bases de datos independientes.

```
% Cargo la base de datos prospectiva:
tabla_nueva = readtable('DatosRecogidos_FILTRADOS.xlsx');
reingresos_nueva = readtable('Reingresos_DatosRecogidos.xlsx');
Y_nueva = reingresos_nueva.Reingreso; % Variable target.

%  Almaceno las 24 primeras variables extraídas de ReliefF (son las
mismas que las aplicadas en el conjunto retrospectivo).
X_nueva = table2array(tabla_nueva(:, 1:24));

% Estandarizo mediante mu y sigma calculados en el modelo original:
X_nueva_estandarizada = (X_nueva - mu) ./ sigma;

% Las predicciones se hacen en este caso sobre el modelo ya ooptimizado.
[~, scores_nueva] = modelo_def.predict(X_nueva_estandarizada);
scores_nueva = scores_nueva(:, 2); % Probabilidad de pertenecer a la
clase positiva.

% Utilizo el umbral óptimo ya calculado:
pred_nueva = double(scores_nueva >= mejor_umbral);

% Matriz de confusión:
matriz_confusion_nueva = confusionmat(Y_nueva, pred_nueva, 'Order', [0
1]);

% Comprobar si a la matriz de confusión le faltan datos:
if sum(sum(matriz_confusion_nueva)) == 0
    error('Matriz de confusión con datos faltantes');
end

TN = matriz_confusion_nueva(1.1); % Verdaderos negativos.
FP = matriz_confusion_nueva(1.2); % Falsos positivos.
FN = matriz_confusion_nueva(2.1); % Falsos negativos.
TP = matriz_confusion_nueva(2.2); % Verdaderos positivos.

% Métricas de rendimiento:
se = TP / (TP + FN); % Sensibilidad.
sp = TN / (TN + FP); % Especificidad.
acc = (TP + TN) / (TP + FN + TN + FP); % Precisión o accuracy.
ppv = TP / (TP + FP); % Valor predictivo positivo.
npv = TN / (TN + FN); % Valor predictivo negativo.
```

```
F1_score_nueva = 2 * ((ppv * se) / (ppv + se)); % F1 score.
LR_pos = se / (1 - sp); % Razón de verosimilitud positiva.
LR_neg = (1 - se) / sp; % Razón de verosimilitud negativa.


fprintf('\n--- RESULTADOS VALIDACIÓN EXTERNA ---\n');
fprintf('Sensibilidad: %.4f\n', se);
fprintf('Especificidad: %.4f\n', sp);
fprintf('Precisión o accuracy: %.4f\n', acc);
fprintf('F1 score: %.4f\n', F1_score_nueva);
fprintf('LR positiva: %.4f\n', LR_pos);
fprintf('LR negativa: %.4f\n', LR_neg);
fprintf('PPV: %.4f\n', ppv);
fprintf('NPV: %.4f\n', npv);
disp('Matriz de confusión:');
disp(matriz_confusion_nueva);

% Obtengo la curva ROC y AUC:
[fpr, tpr, ~, AUC_nueva] = perfcurve(Y_nueva, scores_nueva, 1);
fprintf('AUC del modelo en nuevos datos: %.4f\n', AUC_nueva);

% Represento la curva ROC:
figure;

plot(fpr, tpr, 'LineWidth', 2);
xlabel('1 - Especificidad (FPR)');
ylabel('Sensibilidad (TPR)');
title(sprintf('Curva ROC: Nuevos datos en validación externa (AUC =
%.4f)', AUC_nueva));
grid on;
```

ANEXO 7. Validación temporal prospectiva del modelo predictivo basado en MLP

Se adjunta, análogamente a lo expuesto sobre *Random Forest,* la validación temporal del modelo con la base de datos prospectiva y sobre la red neuronal perceptrón multicapa ya optimizada. Además, se incluyen métricas de rendimiento con las que evaluar su capacidad predictiva en conjuntos de datos nuevos.

```
% Cargo la base de datos prospectiva, que contiene únicamente las
predictoras.
tabla_nueva = readtable('DatosRecogidos_FILTRADOS.xlsx');

% Cargo también la variable target (reingreso):
tabla_reingreso = readtable('Reingresos_DatosRecogidos.xlsx');
Y_nueva_target = tabla_reingreso.Reingreso;
if iscategorical(Y_nueva_target) || isstring(Y_nueva_target) ||
iscellstr(Y_nueva_target)
    Y_nueva_target = double(strcmp(string(Y_nueva_target), 'Sí'));
end

% Almaceno las 24 variables predictoras:
matriz_nueva = table2array(tabla_nueva(:, 1:24));

% Estandarizo con el mismo mu y sigma que en el modelo ya optimizado:
X_nueva_scaled = (matriz_nueva - mu) ./ sigma;

% Hago las predicciones sobre el modelo optimizado obtenido anteriormente
(net final)
Y_nueva_pred_prob = net_final(X_nueva_scaled');
Y_nueva_pred = double(Y_nueva_pred_prob >= umbral_opt);

% Matriz de confusión:
confusion_nueva = confusionmat(Y_nueva_target, Y_nueva_pred);
TN = confusion_nueva(1.1); FP = confusion_nueva(1.2); FN =
confusion_nueva(2.1); TP = confusion_nueva(2.2);

% Muestro los parámetros de la matriz de confusión:
fprintf('Parámetros de matriz de confusión (True Negatives, False
Positives, False Negatives, True Positives): [%d, %d, %d, %d]\n', TN, FP,
FN, TP);

% Métricas de rendimiento manejando NaNs:
if TP + FN > 0
    Se = TP / (TP + FN);
else
    Se = NaN;
    warning('Denominador nulo: TP + FN = 0');
end
```

```
if TN + FP > 0
    Sp = TN / (TN + FP);
else
    Sp = NaN;
    warning('Denominador nulo: TN + FP = 0');
end

if TP + FP > 0
    PPV = TP / (TP + FP);
else
    PPV = NaN;
    warning('Denominador nulo: TP + FP = 0');
end

if TN + FN > 0
    NPV = TN / (TN + FN);
else
    NPV = NaN;
    warning('Denominador nulo: TN + FN = 0');
end

if TP + FN + TN + FP > 0
    Accuracy = (TP + TN) / (TP + FN + TN + FP);
else
    Accuracy = NaN;
    warning('Denominador nulo: TP + FN + TN + FP = 0');
end

if ~isnan(Se) && ~isnan(PPV) && (Se + PPV > 0)
    F1 = 2 * (Se * PPV) / (Se + PPV);
elseif isnan(Se) || isnan(PPV)
    F1 = NaN;
    warning('Sensibilidad o PPV es NaN');
else
    F1 = NaN;
    warning('Denominador nulo: Se + PPV = 0');
end


[X_roc, Y_roc, ~, AUC] = perfcurve(Y_nueva_target, Y_nueva_pred_prob, 1);


fprintf('\n--- Rendimiento de MLP en datos prospectivos (validación
externa) ---\n');
disp('Matriz de confusión:');
disp(confusion_nueva);
fprintf('Sensibilidad: %.3f\n', Se);
fprintf('Especificidad: %.3f\n', Sp);
```

```
fprintf('PPV: %.3f\n', PPV);
fprintf('NPV: %.3f\n', NPV);
fprintf('Precisión o accuracy: %.3f\n', Accuracy);
fprintf('F1 score: %.3f\n', F1);
fprintf('AUC: %.3f\n', AUC);
```

BIBLIOGRAFÍA

[1] "Enfermedad pulmonar obstructiva crónica (EPOC)." Accessed: Jul. 13, 2025. [Online]. Available: https://www.who.int/es/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)

[2] I. Nacional del Corazón and los Pulmones la Sangre, "GUÍA RÁPIDA SOBRE LA ENFERMEDAD PULMONAR OBSTRUCTIVA CRÓNICA," 2023.

[3] R. A. Wise, "MSD_Manuals_fisiopatologiaEPOC," 2024.

[4] "GLOBAL INITIATIVE FOR CHRONIC OBSTRUCTIVE LUNG DISEASE 2025," 2024.

[5] Sociedad Española de Médicos Generales y de Familia, "Hasta un 75% de los casos de EPOC no son identificados en España." Accessed: Aug. 19, 2025. [Online]. Available: https://www.semg.es/index.php/noticias/item/1217-noticia-2024118-2

[6] "Spanish HIE Multimedia - Enfisema." Accessed: Jul. 13, 2025. [Online]. Available: https://ssl.adam.com/content.aspx?productid=118&isarticlelink=false&pid=6&gid=17055&site=nuestrasalud.adam.com&login=NUES7600

[7] A. De Diego Damiá, J. Cortijo Gimeno, M. J. Selma Ferrer, M. León Fabregas, P. Almudever Folch, and J. Milara Paya, "Estudio del efecto de citocinas proinflamatorias en las células epiteliales de pacientes fumadores con o sin EPOC," *Arch Bronconeumol*, vol. 47, no. 9, pp. 447–453, Sep. 2011, doi: 10.1016/J.ARBRES.2011.04.007.

[8] "Spanish HIE Multimedia - Bronquitis." Accessed: Jul. 13, 2025. [Online]. Available: https://ssl.adam.com/content.aspx?productid=118&isarticlelink=false&pid=6&gid=17099&site=nuestrasalud.adam.com&login=NUES7600

[9] O. Rafael Silva, "Fenotipos clínicos en enfermedad pulmonar obstructiva crónica: ¿volver al futuro?," *Rev Med Chil*, vol. 140, no. 7, pp. 926–933, 2012, doi: 10.4067/S0034-98872012000700016.

[10] Asociación de Pacientes con EPOC, "Epidemiología y Fisiopatología de la Enfermedad Pulmonar Obstructiva Crónica (EPOC) - APEPOC - Asociación de Pacientes con EPOC." Accessed: Aug. 18, 2025. [Online]. Available: https://www.apepoc.es/actualidad/68-epidemiologia-y-fisiopatologia-de-la-enfermedad-pulmonar-obstructiva-cronica-epoc

[11] Unidad Técnica de Informe de Estado de Salud y Registros, "IESP - Enfermedad pulmonar obstructiva crónica (EPOC) | Comunidad de Madrid." Accessed: Aug. 18, 2025. [Online]. Available: https://www.comunidad.madrid/servicios/salud/iesp-enfermedad-pulmonar-obstructiva-cronica-epoc

[12] M. Vila, R. Faner, and A. Agustí, "Beyond the COPD-tobacco binomium: New opportunities for the prevention and early treatment of the disease," Jul. 08, 2022, *Ediciones Doyma, S.L*. doi: 10.1016/j.medcli.2022.01.021.

[13] E. García Castillo, G. Vargas, J. A. García Guerra, A. López-Giraldo, and T. Alonso Pérez, "Chronic Obstructive Pulmonary Disease," *Open Respiratory Archives*, vol. 4, no. 2, Apr. 2022, doi: 10.1016/j.opresp.2022.100171.

[14] W. C. Tan *et al.*, "Characteristics of COPD in never-smokers and ever-smokers in the general population: Results from the CanCOLD study," *Thorax*, vol. 70, no. 9, pp. 822–829, Sep. 2015, doi: 10.1136/thoraxjnl-2015-206938.

[15] N. Marchetti *et al.*, "Association between occupational exposure and lung function, respiratory symptoms, and high-resolution computed tomography imaging in COPDGene," *Am J Respir Crit Care Med*, vol. 190, no. 7, pp. 756–762, Oct. 2014, doi: 10.1164/rccm.201403-0493OC.

[16] M. Miravitlles *et al.*, "Spanish COPD Guidelines (GesEPOC) 2021: Updated Pharmacological treatment of stable COPD," *Arch Bronconeumol*, vol. 58, no. 1, pp. 69–81, Jan. 2022, doi: 10.1016/j.arbres.2021.03.005.

[17] J. A. Del Solar H and M. Florenzano V., "ENFERMEDAD PULMONAR OBSTRUCTIVA CRÓNICA."

[18] C. A. Jimenez-Ruiz, J. F. Pascual Lledó, A. Cícero Guerrero, M. Cristóbal Fernández, M. Mayayo Ulibarri, and C. Villar Laguna, "Analysis of quality of life in patients with Chronic Obstructive Pulmonary Disorder (COPD) who give up smoking," *Semergen*, vol. 44, no. 5, pp. 310–315, Jul. 2018, doi: 10.1016/j.semerg.2017.08.003.

[19] GlaxoSmithKline Services Unlimited, "Patient Site Test Page Spanish." Accessed: Aug. 18, 2025. [Online]. Available: https://www.catestonline.org/patient-site-test-page-spainish-spain.html

[20] J. Dreyse, "COPD management in the modern era," *Revista Medica Clinica Las Condes*, vol. 35, no. 3–4, pp. 209–220, May 2024, doi: 10.1016/j.rmclc.2024.05.002.

[21] M. Miravitlles *et al.*, "Spanish COPD Guidelines (GesEPOC): Pharmacological Treatment of Stable COPD," *Arch Bronconeumol*, vol. 48, no. 7, pp. 247–257, 2012, doi: 10.1016/j.arbres.2012.04.001.

[22] A. Cortes Telles, J. A. Cureño Arroyo, A. Elizondo Ríos, R. de J. Hernández Zenteno, and J. Carranza Martínez, "Impacto de las exacerbaciones en la enfermedad pulmonar obstructiva crónica Exacerbaciones en la EPOC," *Respirar*, vol. 15, no. 2, Jun. 2023, doi: 10.55720/respirar.15.2.5.

[23] H. Ruan, H. Zhang, J. Wang, H. Zhao, W. Han, and J. Li, "Readmission rate for acute exacerbation of chronic obstructive pulmonary disease: A systematic review and meta-analysis," Jan. 01, 2023, *W.B. Saunders Ltd*. doi: 10.1016/j.rmed.2022.107090.

[24] J. S. Alqahtani *et al.*, "Risk factors for all-cause hospital readmission following exacerbation of COPD: A systematic review and meta-analysis," *European Respiratory Review*, vol. 29, no. 156, pp. 1–16, 2020, doi: 10.1183/16000617.0166-2019,.

[25] E. Boers *et al.*, "Forecasting the global economic and health burden of chronic obstructive pulmonary disease from 2025 through 2050," *Chest*, vol. 0, no. 0, Apr. 2025, doi: 10.1016/j.chest.2025.03.029.

[26] N. Kapoor, “Why we urgently need to reduce the global burden of COPD | World Economic Forum.” Accessed: Jul. 20, 2025. [Online]. Available: https://www.weforum.org/stories/2024/11/healthcare-reduce-global-burden-copd/

[27] S. Chen *et al.*, “The global economic burden of chronic obstructive pulmonary disease for 204 countries and territories in 2020–50: a health-augmented macroeconomic modelling study,” *Lancet Glob Health*, vol. 11, no. 8, p. e1183, Aug. 2023, doi: 10.1016/S2214-109X(23)00217-6.

[28] “Las agudizaciones graves de la EPOC se asocian a mayores costes sanitarios y mayor mortalidad.” [Online]. Available: https://www.astrazeneca.es/medios/notas-prensa/2022/las_agudizaciones_graves_de_la_EPOC_se_asocian_a_mayores_costes_sanitarios_y_mayor_mortalidad.html#

[29] Ministerio de Sanidad, “Informe Anual del Sistema Nacional de Salud 2023 INFORMES, ESTUDIOS E INVESTIGACIÓN 2025,” 2023.

[30] A. H. Seuc, D. Emma Domínguez, and O. Díaz Díaz, “Instituto Nacional de Endrocrinología INTRODUCCIÓN A LOS DALYS,” 2000.

[31] A. Sanchis, A. Ledezma, J. A. Iglesias, B. García, and J. M. Alonso, “Introducción, Definición y Ejemplos Tipos de MT Equivalencia y Variantes de MT 2.”

[32] “¿Qué es la Inteligencia artificial? Definición, historia y aplicaciones.” Accessed: Jul. 26, 2025. [Online]. Available: https://www.tableau.com/es-mx/data-insights/ai/what-is

[33] T. Lagos Preller, “La generación Z: Konrad Zuse, pionero alemán de la computación - Ministerio Federal de Relaciones Exteriores.” Accessed: Jul. 26, 2025. [Online]. Available: https://alemaniaparati.diplo.de/mxdz-es/aktuelles/konradzuse-1087764

[34] M. Delgado Calvo, “La Inteligencia Artificial. Realidad de un mito moderno,” 1996, Accessed: Jul. 26, 2025. [Online]. Available: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://digibug.ugr.es/bitstream/handle/10481/1270/16912512.pdf?sequence=1

[35] C. Stryker and E. Kavlakoglu, “¿Qué es la Inteligencia Artificial (IA)? | IBM.” Accessed: Jul. 26, 2025. [Online]. Available: https://www.ibm.com/es-es/think/topics/artificial-intelligence

[36] Y. Feng, Y. Wang, C. Zeng, and H. Mao, “Artificial Intelligence and Machine Learning in Chronic Airway Diseases: Focus on Asthma and Chronic Obstructive Pulmonary Disease,” *Int J Med Sci*, vol. 18, no. 13, pp. 2871–2889, 2021, doi: 10.7150/IJMS.58191.

[37] J. Holdsworth and M. Scapicchio, “¿Qué es el deep learning? | IBM.” Accessed: Jul. 26, 2025. [Online]. Available: https://www.ibm.com/es-es/think/topics/deep-learning

[38] C. Stryker and M. Scapicchio, “¿Qué es la IA generativa? | IBM.” Accessed: Jul. 26, 2025. [Online]. Available: https://www.ibm.com/es-es/think/topics/generative-ai

[39] A. Agusti and M. Vila, “Artificial Intelligence in COPD,” May 01, 2025, *Sociedad Espanola de Neumologia y Cirugia Toracica (SEPAR)*. doi: 10.1016/j.arbres.2024.12.013.

[40] “Inteligencia Artificial y EPOC: Un Futuro de Oportunidades - EPOC España.” Accessed: Jul. 26, 2025. [Online]. Available: https://www.epocespana.org/2025/02/20/inteligencia-artificial-y-epoc-un-futuro-de-oportunidades/

[41] L. Vargas-Ramirez and R. Brango Ayazo, “Artículo de Revisión INTELIGENCIA ARTIFICIAL EN NEUMOLOGÍA,” Jan. 2022. [Online]. Available: https://www.researchgate.net/publication/358040724

[42] B. Zhang *et al.*, “Machine learning in chronic obstructive pulmonary disease,” *Chin Med J (Engl)*, vol. 136, no. 5, pp. 536–538, Mar. 2023, doi: 10.1097/CM9.0000000000002247.

[43] Z. Xu, F. Li, Y. Xin, Y. Wang, and Y. Wang, “Prognostic risk prediction model for patients with acute exacerbation of chronic obstructive pulmonary disease (AECOPD): a systematic review and meta-analysis,” Dec. 01, 2024, *BioMed Central Ltd*. doi: 10.1186/s12931-024-03033-4.

[44] J. Cid Ruzafa and J. Damián Moreno, “Valoración de la discapacidad física: el indice de Barthel,” 1997. Accessed: Aug. 20, 2025. [Online]. Available: https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1135-57271997000200004

[45] M. Herdman, X. Badia, and S. Berra, “El EuroQol-5D: una alternativa sencilla para la medición de la calidad de vida relacionada con la salud en atención primaria | Atención Primaria.” Accessed: Aug. 20, 2025. [Online]. Available: https://www.elsevier.es/es-revista-atencion-primaria-27-articulo-el-euroqol-5d-una-alternativa-sencilla-13020211

[46] Comité Científico del Proyecto TAI y Chiesi, “Test de Adhesión a los Inhaladores.” Accessed: Aug. 20, 2025. [Online]. Available: https://www.taitest.com/

[47] M. Charlson, “Charlson Comorbidity Index (CCI).” Accessed: Jul. 30, 2025. [Online]. Available: https://www.mdcalc.com/calc/3917/charlson-comorbidity-index-cci

[48] H. Kang, “The prevention and handling of the missing data,” *Korean J Anesthesiol*, vol. 64, no. 5, p. 402, May 2013, doi: 10.4097/KJAE.2013.64.5.402.

[49] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, “A survey on missing data in machine learning,” *J Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00516-9.

[50] M. J. Bernal Vélez, I. M. Henao, M. Isabel, and I. Arango, “Predicción de enfermedades del corazón usando el algoritmo K-Nearest Neighbors,” 2022. [Online]. Available: https://www.researchgate.net/publication/364476395

[51] S. Zhang, “Nearest neighbor selection for iteratively kNN imputation,” *Journal of Systems and Software*, vol. 85, no. 11, pp. 2541–2552, Nov. 2012, doi: 10.1016/j.jss.2012.05.073.

[52] S. G. Liao *et al.*, "Missing value imputation in high-dimensional phenomic data: Imputable or not, and how?," *BMC Bioinformatics*, vol. 15, no. 1, Nov. 2014, doi: 10.1186/s12859-014-0346-6.

[53] A. Sania *et al.*, "K-nearest neighbor algorithm for imputing missing longitudinal prenatal alcohol data," *Advances in Drug and Alcohol Research*, vol. 4, 2024, doi: 10.3389/adar.2024.13449.

[54] T. Pukkala, N. Aquilué, A. Just, J. Corbera, and A. Trasobares, "Developing kNN forest data imputation for Catalonia," *J For Res (Harbin)*, vol. 35, no. 1, Dec. 2024, doi: 10.1007/s11676-024-01735-5.

[55] L. Beretta and A. Santaniello, "Nearest neighbor imputation algorithms: A critical evaluation," *BMC Med Inform Decis Mak*, vol. 16, Jul. 2016, doi: 10.1186/s12911-016-0318-z.

[56] N. Jain, "¿Qué es el análisis de datos en la sanidad? Definición, importancia, ejemplos, beneficios y análisis de Big Data." Accessed: Jul. 31, 2025. [Online]. Available: https://ideascale.com/es/blogs/que-es-el-analisis-datos-en-la-sanidad/

[57] P. Faraldo and B. Pateiro, "Estadística y metodología de la investigación Tema 1. Estadística Descriptiva," 2012, Accessed: Jul. 31, 2025. [Online]. Available: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_G2021103104_EstadisticaTema1.pdf

[58] R. M. Álvarez Esteban, "ESTADÍSTICA DESCRIPTIVA." [Online]. Available: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.cartagena99.com/recursos/alumnos/apuntes/T3%20completo.pdf

[59] G. Westreicher, "Rango Intercuartílico (RIC) ¿Qué es? Definición | Rankia." Accessed: Aug. 01, 2025. [Online]. Available: https://www.rankia.com/diccionario/economia/rango-intercuartilico-ric

[60] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," Sep. 01, 2018, *Academic Press Inc.* doi: 10.1016/j.jbi.2018.07.014.

[61] R. J. Urbanowicz, R. S. Olson, P. Schmitt, M. Meeker, and J. H. Moore, "Benchmarking relief-based feature selection methods for bioinformatics data mining," *J Biomed Inform*, vol. 85, pp. 168–188, Sep. 2018, doi: 10.1016/j.jbi.2018.07.015.

[62] J. J. Espinosa Zúñiga, "Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito," *Ingeniería Investigación y Tecnología*, vol. 21, no. 3, pp. 1–16, Jul. 2020, doi: 10.22201/fi.25940732e.2020.21.3.022.

[63] J. Redondo Antón, "Comparativa de modelos de random forest y redes neuronales aplicados al mantenimiento predictivo con valores ausentes y datos desbalanceados," Universidad Complutense de Madrid, Madrid, 2021.

[64] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/bjml/2024/007.

[65] "TreeBagger - Ensemble of bagged decision trees - MATLAB." Accessed: Aug. 18, 2025. [Online]. Available: https://es.mathworks.com/help/stats/treebagger.html

[66] I. M. Galván and J. M. Valls, "Perceptrón Multicapa. REDES DE NEURONAS ARTIFICIALES," Madrid. [Online]. Available: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://ocw.uc3m.es/pluginfile.php/2241/mod_page/content/38/Transparencias_tema3.pdf

[67] A. Palmer, J. J. Montaño, and R. Jiménez, "Tutorial sobre Redes Neuronales Artificiales: El Perceptrón Multicapa," 2001. [Online]. Available: https://www.researchgate.net/publication/331640802

[68] M. Yepes, "RPubs - Funciones de Activación." Accessed: Aug. 18, 2025. [Online]. Available: https://rpubs.com/mariangely/1333826

[69] M. Rafly Indrawan, "Understanding Multilayer Perceptron: The Foundation of Modern Neural Networks | by Muhammad Rafly Indrawan | Medium." Accessed: Aug. 18, 2025. [Online]. Available: https://medium.com/@muhammadraflyindrawan/understanding-multilayer-perceptron-the-foundation-of-modern-neural-networks-5b8bf757db99

[70] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," *African Journal of Biomedical Research*, pp. 4023–4031, Nov. 2024, doi: 10.53555/ajbr.v27i4s.4345.

[71] J. H. Cabot and E. G. Ross, "Evaluating Prediction Model Performance," *Surgery*, vol. 174, no. 3, p. 723, Sep. 2023, doi: 10.1016/J.SURG.2023.05.023.

[72] B. J. Erickson and F. Kitamura, "Magician's Corner: 9. Performance Metrics for Machine Learning Models," *Radiol Artif Intell*, vol. 3, no. 3, p. e200126, May 2021, doi: 10.1148/RYAI.2021200126.

[73] S. Bravo-Grau and J. P. Cruz Q., "Estudios de exactitud diagnóstica: Herramientas para su Interpretación," *Revista chilena de radiología*, vol. 21, no. 4, pp. 158–164, 2015, doi: 10.4067/S0717-93082015000400007.

[74] DATAtab, "Prueba U de Mann-Whitney - Explicación sencilla - DATAtab." Accessed: Aug. 19, 2025. [Online]. Available: https://datatab.es/tutorial/mann-whitney-u-test

[75] S. Siegel and N. J. Castellan, *Estadística no paramétrica aplicada a las ciencias de la conducta*. 1995.

[76] John Wiley & Sons, "MANN-WHITNEY U TEST." 2009. Accessed: Aug. 19, 2025. [Online]. Available: https://sci-hub.red/https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470479216.corpsy0524

[77] J. Frost, "Fisher's Exact Test: Using & Interpreting - Statistics By Jim." Accessed: Aug. 19, 2025. [Online]. Available: https://statisticsbyjim.com/hypothesis-testing/fishers-exact-test/

[78] J. Amat Rodrigo, "Test estadísticos para variables cualitativas: Test exacto de Fisher, chi-cuadrado de Pearson, McNemar y Q-Cochran." Accessed: Aug. 19, 2025. [Online]. Available:

https://cienciadedatos.net/documentos/22.2_test_exacto_de_fisher_chi-cuadrado_de_pearson_mcnemar_qcochran

[79] A. Vierra, A. Razzaq, and A. Andreadis, "Categorical Variable Analyses: Chi-square, Fisher Exact, and Mantel–Haenszel," *Handbook for Designing and Conducting Clinical and Translational Surgery*, pp. 171–175, Jan. 2023, doi: 10.1016/B978-0-323-90300-4.00095-1.

[80] J. de Miguel Díez, T. G. García, and L. P. Maestu, "Comorbilidades de la EPOC," *Arch Bronconeumol*, vol. 46, no. SUPPL.11, pp. 20–25, Jan. 2010, doi: 10.1016/S0300-2896(10)70058-2.

[81] R. Vargas García, "Estudio de las arritmias cardíacas en la enfermedad pulmonar obstructiva," 2002, Accessed: Aug. 13, 2025. [Online]. Available: https://dialnet.unirioja.es/servlet/tesis?codigo=233943&info=resumen&idioma=SPA

[82] "Gasometría arterial: MedlinePlus enciclopedia médica." Accessed: Aug. 14, 2025. [Online]. Available: https://medlineplus.gov/spanish/ency/article/003855.htm

[83] B. S. Cabral, D. M. Castro, and S. Fabbiani, "Lugar en la terapéutica de los mucolíticos en la Enfermedad Pulmonar Obstructiva Crónica (EPOC)."

[84] "Ventilación no invasiva en pacientes con exacerbacion de EPOC." Accessed: Aug. 14, 2025. [Online]. Available: https://www.scielo.org.ar/scielo.php?pid=S0025-76802007000200002&script=sci_arttext&tlng=en

[85] N. Horita, N. Miyazawa, R. Kojima, M. Inoue, Y. Ishigatsubo, and T. Kaneko, "Uso crónico de teofilina y mortalidad en la enfermedad pulmonar obstructiva crónica: un metaanálisis," *Arch Bronconeumol*, vol. 52, no. 5, pp. 233–238, May 2016, doi: 10.1016/J.ARBRES.2015.02.021.

[86] K. Portillo Carroz, "La anemia en la EPOC. ¿Debemos pensar en ello?," *Arch Bronconeumol*, vol. 43, no. 7, pp. 392–398, Jul. 2007, doi: 10.1157/13107696.

[87] F. López García, M. Pineda Cuenca, and J. Custardoy Olavarrieta, "10-comorbilidad-epoc-ansiedad-depresion".

[88] "Agudización de la EPOC," *Arch Bronconeumol*, vol. 53, pp. 46–62, Jun. 2017, doi: 10.1016/S0300-2896(17)30369-1.

[89] J. Marín Trigo, "Principales parámetros de función pulmonar en la enfermedad pulmonar obstructiva crónica (EPOC) | Atención Primaria." Accessed: Aug. 14, 2025. [Online]. Available: https://www.elsevier.es/es-revista-atencion-primaria-27-articulo-principales-parametros-funcion-pulmonar-enfermedad-pulmonar-obstructiva-13049899

[90] C. Y. Wu, C. N. Hsu, C. Wang, J. Y. Chien, C. C. Wang, and F. J. Lin, "Predicting outcomes after hospitalisation for COPD exacerbation using machine learning," *ERJ Open Res*, vol. 11, no. 3, May 2025, doi: 10.1183/23120541.00651-2024.

[91] L. Chen and S. Chen, "Prediction of readmission in patients with acute exacerbation of chronic obstructive pulmonary disease within one year after treatment and

discharge," *BMC Pulm Med*, vol. 21, no. 1, Dec. 2021, doi: 10.1186/s12890-021-01692-3.

[92] K. Lopez *et al.*, "Deep learning prediction of hospital readmissions for asthma and COPD," *Respir Res*, vol. 24, no. 1, Dec. 2023, doi: 10.1186/s12931-023-02628-7.