ULPGC
Universidad de
Las Palmas de
Gran Canaria

ESCUELA DE INGENIERÍA DE TELECOMUNICACIÓN Y ELECTRÓNICA

TRABAJO DE FIN DE MÁSTER

Diseño de un método de monitorización de pacientes con sepsis en Unidades de Cuidados Intensivos mediante técnicas de Machine Learning

Titulación: Máster Universitario en Ingeniería de Telecomunicación

Autora: Caterina Ríos Bolaños

Tutores: Dr. Juan Luis Navarro Mesa, Francisco J. Suárez Díaz, Encarnación Gimeno Nieves

Fecha: 14 de julio de 2022

Resumen

La sepsis es una afección que se produce como consecuencia de una respuesta extrema del cuerpo frente a una infección que daña sus propios órganos y tejidos, causando secuelas irreversibles e incluso la muerte. En la actualidad, representa una crisis sanitaria a nivel mundial que afecta a millones de personas. Los métodos de evaluación tradicionales no permiten el diagnóstico precoz de la afección, que ha demostrado ser unos de los principales factores a la hora de reducir la mortalidad de los pacientes. Estudios recientes demuestran que la combinación de datos clínicos con técnicas de Machine Learning permite predecir la sepsis con precisión varias horas antes de que los síntomas físicos se manifiesten.

En la línea de los descrito anteriormente, el objetivo principal de este proyecto es el diseño de un método de detección temprana de la sepsis que sirva como apoyo a los médicos en las Unidades de Cuidados Intensivos. Para lograrlo, se ha hecho un diseño que integra clasificadores convencionales obtenidos a partir de las variables clínicas (es lo que se ha denominado "clasificadores base"), estrategias de combinación de clasificadores, y un elemento de pos-procesado basado en probabilidades acumuladas en ventanas temporales deslizantes. El desarrollo es de tipo constructivo en el que cada nuevo elemento añade una mejora en términos de detección de sepsis. Los clasificadores base que se han usado son Bayesiano, SVM, y regresión lineal y logística. Las estrategias de combinación han sido promedio, producto y *majority voting*.

A partir del diseño anterior, se han diseñado dos módulos: uno de identificación temprana de pacientes sépticos y otro de detección temprana de eventos de sepsis. El primero permite identificar en las primeras 8 horas de ingreso en la UCI qué pacientes podrían llegar a desarrollar una sepsis. Por otro lado, el segundo permite distinguir cuándo un paciente pasa de no tener sepsis a tenerla, con especial énfasis en las 6 horas anteriores al comienzo.

Para probar las bondades del diseño se ha llevado a cabo una experimentación intensiva sobre la base de datos aportada en el "Physionet Challenge: Early Prediction of Sepsis from Clinical Data. The PhysioNet Computing in Cardiology Challenge, 2019". En él se aportan 40.300 pacientes provenientes de dos hospitales, divididos en los conjuntos A (20.300) y B (20.000). En cada conjunto hay pacientes con y sin sepsis, y de ellos hay datos clínicos que permiten abordar el problema de detección temprana. Así mismo, en

el reto se introduce una “función de utilidad” que expresa la expectativa que los médicos tienen acerca de los resultados de un método de monitorización, y permite cuantificar la bondad de los métodos desarrollados.

De todos los clasificadores base aplicados, el Bayesiano ha resultado ser el mejor. A su vez, las estrategias de combinación de tipo producto y *majority voting* han resultado las mejores. La ventana deslizante que mejores resultados ha dado para las probabilidades acumuladas es de 8 horas.

Como síntesis del resultado de todo el diseño, se puede resaltar que los valores de utilidad son 0,675 para el conjunto de datos A, 0,641 para el conjunto de datos B y 0,621 para la combinación de ambos conjuntos. Estos resultados, y otros que se muestran en la memoria, permiten asegurar que el sistema propuesto es altamente competitivo en el marco científico de referencia en el que se ubica.

Abstract

Sepsis is a condition that occurs when the body has an extreme response to an infection which damages its own organs and tissues, leading to irreversible sequels and even death. As of now, it is a world-wide health crisis that affects millions of people every year. Traditional evaluation methods do not allow the early diagnosis of sepsis, which has been proved to be a key factor to reduce mortality in patients. Studies show that combining medical data with Machine Learning techniques allows to accurately predict sepsis several hours before the physical symptoms manifest themselves.

In line with the above, the main goal of this study is the development of a sepsis early detection system capable of providing support for healthcare professionals in the ICUs. To achieve this, a design has been made that integrates conventional classifiers derived from clinical variables (referred to as "base classifiers"), classifier combination strategies, and a post-processing element based on cumulative probabilities in sliding time windows. The development is constructive in that each new element adds an improvement in terms of sepsis detection. The base classifiers that have been used are Bayesian, SVM, and linear and logistic regression. The combination strategies used have been average, product and majority voting.

Based on the above design, two modules have been designed: one for early identification of septic patients and the other for early detection of sepsis events. The first one allows to identify within the first 8 hours of admission to the ICU which patients could develop sepsis. On the other hand, the second allows to distinguish when a patient goes from not having sepsis to having sepsis, with special emphasis on the 6 hours prior to the onset of sepsis.

To test the design, intensive experimentation has been carried out on the database provided in the "Physionet Challenge: Early Prediction of Sepsis from Clinical Data. The PhysioNet Computing in Cardiology Challenge, 2019". It provides 40,300 patients from two hospitals, divided into sets A (20,300) and B (20,000). In each set there are patients with and without sepsis, and from them there are clinical data to address the problem of early detection. The challenge also introduces a "utility function" that expresses the expectation that physicians have about the results of a monitoring method and allows quantifying the goodness of the developed methods.

Of all the base classifiers applied, the Bayesian classifier proved to be the best. In turn, the product and majority voting combination strategies were the best. The sliding window that has given the best results for the cumulative probabilities is 8 hours.

As a conclusion of the result of the whole design, it can be highlighted that the utility values are 0.675 for data set A, 0.641 for data set B and 0.621 for the combination of both sets. These results, and others shown in the report, allow us to ensure that the proposed system is highly competitive in the scientific reference framework in which it is located.

Índice de contenido

Índice de figuras

Índice de tablas

Glosario de acrónimos

APACHE: Acute Physiology and Chronic Health Evaluation.

AUROC: Area Under Receiver Operating Characteristic Curve.

BISEPRO: Big data Sepsis PROject.

FN: False Negative.

FP: False Positive.

FVN: Fracción de Verdaderos Negativos.

FVP: Fracción de Verdaderos Positivos.

HMM: Hidden Markov Models.

ICULOS: ICU length-of-stay.

IQR: Interquartile Range.

MAP: Maximum a Posteriori.

MPM: Mortality Probability Models.

PCA: Principal Component Analysis.

PCM: Pulse Code Modulation.

qSOFA: Quick SOFA.

ROC: Receiver Operating Characteristic.

SAPS*:* Simplified Acute Physiology Score.

SIRS: Systemic Inflammatory Response Syndrome.

SOFA: Sequential Organ Failure Assessment.

SVM: Support Vector Machines.

TN: True Negative.

TP: True Positive.

UCI: Unidad de Cuidados Intensivos.

VPN: Valor Predictivo Negativo.

VPP: Valor Predictivo Positivo.

Parte I: Memoria

Capítulo 1. Introducción

1.1. Antecedentes

La sepsis se define formalmente como una disfunción orgánica potencialmente mortal causada por una respuesta desregulada del huésped a una infección. El shock séptico es un subconjunto de la sepsis en el que las anomalías circulatorias, celulares y metabólicas subyacentes son lo suficientemente importantes como para aumentar la probabilidad de muerte de un paciente de forma sustancial [1]. Dicho de otra forma, la sepsis es una afección que se produce cuando el cuerpo sufre una respuesta violenta y descontrolada a una infección, lo que provoca que sus propios órganos y tejidos resulten dañados. La sepsis puede causar secuelas irreversibles en el paciente, e incluso la muerte.

La aparición de sepsis suele estar relacionada con microorganismos, aunque también puede producirse debido a incidentes no infecciosos como traumatismos graves, neumonía, pancreatitis e infecciones del sistema urinario, entre otros. A nivel físico, la sepsis suele manifestarse en forma de fiebre, inestabilidad mental, hipotensión temporal, disminución de la cantidad de orina o trombocitopenia (falta de plaquetas en la sangre) inexplicable. Si no se adoptan las medidas necesarias para tratar la enfermedad a tiempo, el paciente puede llegar a sufrir insuficiencia respiratoria y renal, trastornos de la coagulación e hipotensión, todos trastornos potencialmente mortales [2].

La sepsis es una crisis sanitaria de nivel mundial. La *Global Sepsis Alliance* calcula que afecta a entre 47 y 50 millones de personas cada año, de las cuales al menos 11 millones acaban muriendo. Esto implica una muerte cada 2,8 segundos de media [3]. Por otro lado, la Organización Mundial de la Salud (OMS) alerta de que en el año 2017 las defunciones relacionadas con la sepsis representaron casi el 20% de todas las muertes a nivel global. De este porcentaje, al menos 2.9 millones eran niños menores de 5 años. También se aprecia disparidad en la incidencia de la sepsis a nivel regional, ya que aproximadamente el 85% de los casos y las muertes asociadas a la sepsis se dieron en países con ingresos bajos y medios [4].

Además de las altas tasas de mortalidad, la gestión de la sepsis supone un importante gasto económico al sistema sanitario. Por ejemplo, en Estados Unidos, el coste del tratamiento de la sepsis en los hospitales supuso más de 24.000 millones de dólares en el

año 2013, lo que representa el 13% del total de los costes sanitarios del país. Sin embargo, los casos de sepsis solo representaron el 3,6% de las estancias hospitalarias [5].

En España se declaran unos 50.000 casos de sepsis cada año, aunque se estima que la cifra real podría ser hasta 5 veces mayor. Se calcula que unas 17.000 personas mueren al año en España debido a esta afección [6]. Con el objetivo de reducir el riesgo en los pacientes, se han llevado a cabo diferentes iniciativas en los hospitales nacionales. Por ejemplo, el Código Sepsis es un conjunto de alertas automatizados que ayudan a identificar escenarios de sepsis potencialmente graves. Combinado con técnicas de inteligencia artificial puede conseguir incluso la clasificación de los pacientes en función de la probabilidad que tengan de sufrir sepsis. En las Unidades de Cuidados Intensivos (UCI) donde se ha implantado se ha conseguido reducir la mortalidad hasta por debajo del 20% [7].

Un proyecto a destacar es el *Big data Sepsis PROject* (BISEPRO) diseñado por la Unidad de Sepsis del Hospital Universitario Son Llàtzer de Palma de Mallorca [8]. Este *software* analiza en tiempo real los datos de los pacientes, que incluyen información clínica, analítica, farmacológica, microbiológica y de antecedentes personales. La herramienta actualiza cada 15-30 minutos la evaluación de todos los pacientes que son atendidos en cualquier departamento del hospital, y cada vez que detecte un posible caso de sepsis, envía al médico una alerta que se asocia a una luz amarilla, naranja o roja en función de la certeza. De esta manera, el equipo multidisciplinar es capaz de valorar y confirmar el diagnóstico [9].

La identificación precoz de la sepsis es, por tanto, un aspecto fundamental para reducir las altas tasas de ingreso y mortalidad de pacientes en las UCI. En los últimos años se ha investigado la aplicación de técnicas de inteligencia artificial a los datos recogidos en los registros médicos electrónicos como método de detección temprana de la enfermedad. Por ejemplo, en [10] se presenta un modelo que emplea el aprendizaje profundo sobre un conjunto de datos de múltiples centros hospitalarios daneses. El modelo está compuesto por la combinación de redes neuronales convolucionales y una red *Long Short-Term Memory* (LSTM). Los resultados mostraron que usando este método se podía detectar el inicio de la sepsis entre 3 y 24 horas antes.

Otro enfoque distinto es el descrito en [11]. En este caso, se intenta abordar el problema del inicio tardío de sepsis neonatal. En concreto, se analiza un método para detectar la

aparición de sepsis sin necesidad de tomar una muestra de sangre al recién nacido. El método consiste en la observación y modelado de eventos fisiológicos en tiempo real empleando un Modelo Oculto de Markov Autorregresivo (*Autorregressive Hiden Markov Model*, AR-HMM).

Este proyecto parte del trabajo realizado por Alba Manso [12] bajo la codirección de dos de mis tutores, y pretende explorar las líneas abiertas que se definieron en el mismo. Es necesario precisar que estas líneas han dado lugar a los objetivos que se han marcado, así como otros que han surgido del planteamiento y en el devenir del proyecto. El algoritmo de detección desarrollado en [12] se entrenó y se probó sobre el conjunto de datos de libre acceso proporcionados por PhysioNet [13] para abordar el reto 2019 de predicción temprana de sepsis [14]. Este conjunto de datos está formado por un total de 40.336 registros de pacientes de dos sistemas hospitalarios distintos, y contiene muestras clasificadas como sepsis o no sepsis, hora a hora, por lo que constituye un valioso recurso a la hora de entrenar y validar los algoritmos.

1.2. Objetivos

El objetivo principal de este proyecto es desarrollar un algoritmo que facilite la detección temprana de la sepsis a partir de un conjunto de datos (*dataset*) clínicos, y en la medida de lo posible predecir su aparición.

Para ello, en primer lugar, se estudiará la funcionalidad del algoritmo desarrollado para incluir explícitamente la evolución temporal de cada paciente. Por ejemplo, en [12] se proponen los Modelos Ocultos de Markov (*Hidden Markov Models*, HMM) y las redes bayesianas en general como una alternativa de interés, aunque se estudiarán otros métodos que puedan ser adecuados para el modelo a desarrollar. En este caso, se ha partido de un planteamiento en el que todas las opciones de Aprendizaje Automático estaban abiertas y se han buscado las mejores opciones científico-técnicas para conseguir una buena detección temprana.

En segundo lugar, en caso de necesidad, se estudiará el uso del aprendizaje profundo aplicado a la problemática de la sepsis. Las redes neuronales profundas son una opción a tener en cuenta para este objetivo. El resultado de este estudio sería decidir qué tipo de algoritmo utilizar de entre los consultados (HMM, LSTM, etc.) o plantear una aportación propia.

Por último, en caso de ser posible, se buscarán nuevos conjuntos de datos clínicos que aporten mayor riqueza de datos, como pueden ser un período de muestreo inferior a una hora, de forma que se disponga de datos de mejor calidad y se reduzca la tasa de datos perdidos.

En definitiva, el objetivo general se puede desglosar en los siguientes objetivos específicos:

- O1: Ampliar la funcionalidad del algoritmo de detección temprana para mostrar la evolución temporal de los pacientes.
- O2: Estudiar técnicas Machine Learning y, de ser necesario, de aprendizaje profundo que puedan incorporarse al algoritmo.

1.3. Peticionario

Este Trabajo Fin de Máster ha sido elaborado como parte de los estudios que conllevan a la obtención del Máster Universitario en Ingeniería de Telecomunicación en la Escuela de Ingeniería de Telecomunicación y Electrónica de la Universidad de Las Palmas de Gran Canaria.

1.4. Estructura del documento

El presente documento está divido en diferentes capítulos. El primer capítulo consiste en una introducción al trabajo propuesto y los antecedentes que han llevado a su desarrollo, así como los objetivos marcados para este TFM.

En el segundo capítulo se analiza el marco teórico en el que se ubica este trabajo. Se hace una revisión de las diferentes definiciones de la sepsis que se han formulado a lo largo de la historia, así como de las distintas escalas pronósticas que se han utilizado tradicionalmente para el diagnóstico y la evaluación de esta enfermedad. El capítulo finaliza con una pequeña introducción a la Inteligencia Artificial aplicada a esta problemática.

El tercer capítulo trata el marco metodológico en el que se enmarca el proyecto. En este capítulo de hace una revisión exhaustiva de todas las tecnologías que se han empleado durante el desarrollo de este TFM. De entre ellas, se puede destacar las distintas técnicas de Machine Learning analizadas, el teorema de Bayes, la probabilidad de detección acumulada y las diferentes métricas de calidad con las que se ha llevado a cabo la evaluación del sistema propuesto.

El cuarto capítulo está dedicado al análisis de los datos con los que se ha trabajado. Por un lado, se ha tratado la estructura general de los datos y sus características generales, incluyendo la prevalencia de las clases y las variables clínicas disponibles. Por otro lado, se ha hecho un estudio sobre la cantidad de datos perdidos y de valores atípicos y aberrantes en el conjunto de datos. Asimismo, se ha analizado el impacto de éstos de cara al proyecto.

En el quinto capítulo se describe en profundidad el sistema de monitorización propuesto. Primero, se han explicado los criterios de selección de pacientes y variables clínicas. Luego, se ha explicado detalladamente el funcionamiento del sistema a través de los diferentes módulos que lo conforman. Finalmente, se ha descrito la metodología de evaluación del sistema y se han presentado los resultados obtenidos. Este capítulo incluye, además, una compativa de las diferentes técnicas básicas de Machine Learning que se han probado.

El sexto y último capítulo está dedicado a las conclusiones del trabajo. En este capítulo se ha valorado si se han cumplido los objetivos marcados al comienzo del proyecto y se analiza la bondad de los resultados obtenidos. Asimismo, se proponen posibles líneas de trabajo futuras.

Capítulo 2. Marco teórico general

A lo largo de la historia, la sepsis ha sido una de las principales causas de mortalidad a nivel mundial. Actualmente es, además, una de las afecciones que más gastos hospitalarios conlleva y que más recursos clínicos consume. A pesar de que la sepsis ha sido conocida en una u otra forma desde hace miles de años, los síntomas que la identifican y su tratamiento han variado a lo largo del tiempo conforme la comprensión de dicha enfermedad ha avanzado. En este capítulo se trata el concepto de la sepsis y los diferentes sistemas de evaluación y tratamiento de la afección que existen actualmente.

2.1. La sepsis

El término "sepsis" proviene del vocablo griego *seps*, que significa "descomposición de materia orgánica animal o vegetal". Hipócrates es el primero en usar esta palabra en la literatura médica en los *Tratados hipocráticos*, donde hacía alusión a un concepto egipcio que relaciona la aparición de algunas enfermedades con la intoxicación por el consumo de productos perjudiciales. A partir del trabajo de Hipócrates, Galeno introdujo en el siglo II el concepto *"Pus bonum et laudabile",* es decir, "Pus buena y digna de alabanza", estableciendo de esta forma que era necesario la infección de la herida antes de que ésta pudiera cicatrizar. Este enfoque estimuló, durante la Edad Media, el uso indiscriminado del cauterio, así como el uso de ungüentos compuestos por sustancias putrefactas o cáusticas para facilitar la supuración en la lesión. Las primeras críticas al trabajo de Galeno aparecieron en el siglo XIII, cuando se empezaron a proponer nuevos tratamientos que evitaran la aparición de pus. No obstante, estos trabajos fueron duramente criticados, y no fue hasta el Renacimiento que las practicas comenzaron a cambiar. Finalmente, a finales del siglo XIX, importantes científicos como Pasteur, Koch o Lister demostraron la existencia de los microorganismos y su relación con la infección de las heridas [15].

Debido a la compleja naturaleza de la afección y a los constantes avances médicos, la definición de sepsis ha ido variando a lo largo de los años. Actualmente, la definición aceptada a nivel mundial es la conocida como Sepsis-3.

2.1.1. Sepsis-1

En el año 1992 se celebró el primer consenso para acordar una definición universal de la sepsis, así como determinar unos criterios de diagnóstico que permitieran hacer diagnosticar la enfermedad de manera precoz. En este consenso se definió la sepsis como

la “respuesta inflamatoria sistémica asociada a una infección”. Del mismo modo, se definió el concepto “sepsis severa” como la aparición de disfunción orgánica, hipotensión arterial o hipotensión tisular persistente asociadas a la sepsis. Un caso de sepsis severa que persiste a pesar de la administración y reanimación adecuada de líquidos podía convertirse en un “*shock* séptico” [16] .

Como medio de diagnóstico, se propuso el uso del Síndrome de Respuesta Inflamatoria Sistémica (*Systemic Inflammatory Response Syndrome*, SIRS). Los criterios SIRS se explican más detalladamente en el apartado 2.2.1.

2.1.2. Sepsis-2

En el año 2001 se decidió que era necesario revisar el concepto de sepsis para reflejar el nuevo conocimiento adquirido acerca de la enfermedad. Para alcanzar este objetivo, representantes de diversas asociaciones norteamericanas y europeas celebraron una conferencia en la que evaluaron diferentes áreas: signos y síntomas de sepsis, marcadores celulares, citoquinas, datos microbiológicos y parámetros de coagulación. Como resultado de esta conferencia, se determinó que la definición Sepsis-1 seguía vigente, y se expandió la lista de signos y síntomas relacionados con las sepsis para reflejar con mayor precisión la respuesta clínica a la infección [17].

Por otro lado, dado que la sepsis es una enfermedad compleja que afecta a cada individuo de forma distinta, se consideró necesario adoptar un nuevo enfoque para realizar su diagnóstico. Con este objetivo en mente, se introdujo el sistema PIRO. Este sistema se basa en cuatro conceptos diferenciados [18]:

- La Predisposición (P) a sufrir la enfermedad, lo que afecta en gran medida a que el paciente tenga una evolución desfavorable. Múltiples factores como la genética, la edad o patologías subyacentes pueden influir en la predisposición del paciente.
- La Infección (I), ya que la evolución del paciente puede verse influida en mayor o menor medida dependiendo del agente infeccioso, la zona afectada y su extensión.
- La Respuesta (R) del organismo a la infección, en la que se liberan mediadores de la inflamación, condicionará la aparición de SIRS. Si se produce un desbalance entre los diferentes agentes mediadores, puede generarse una reacción inflamatoria potencialmente mortal.

- La Disfunción Orgánica (O) es el factor principal que determina la evolución del paciente. El esfuerzo terapéutico y el consumo de recursos clínicos son proporcionales al número de órganos disfuncionales. Asimismo, la evolución de éstos es el mayor indicador para conocer si el paciente se está recuperando o deteriorándose.

Debido a la gran complejidad que presenta, la utilización del sistema PIRO no se ha extendido. En su lugar, se emplean distintas escalas pronósticas [19] que permiten evaluar de forma rápida y sencilla la enfermedad en un paciente.

2.1.3. Sepsis-3

En el año 2017, los importantes avances en los campos de patología, gestión y epidemiología de la sepsis motivaron una nueva revisión de la definición formal de esta afección. Identificaron varias limitaciones en las definiciones anteriores, de entre las que destacaban un enfoque excesivo en la inflamación, la noción errónea de que la sepsis grave evoluciona de forma continua hacia el *shock* séptico, y una especificidad y sensibilidad inadecuadas de los criterios SIRS. Asimismo, se concluyó que el término "sepsis grave" era redundante, por lo que fue eliminado.

Las definiciones alcanzadas por el grupo de expertos se conocen como Sepsis-3, y son las siguientes:

- La sepsis es una disfunción orgánica potencialmente mortal causada por una respuesta desregulada del huésped a la infección.
- La disfunción orgánica está representada por un aumento de dos puntos o más en la escala de la Evaluación de Fallo Orgánico Secuencial (SOFA, *Sequential Organ Failure Assessment*). Esta puntuación está asociada con una mortalidad superior al 10%. En pacientes en los que desconoce previamente si sufren o no disfunción orgánica, la puntuación SOFA de partida puede considerarse cero.
- El shock séptico es un subconjunto de la sepsis en el que las anomalías circulatorias y celulares o metabólicas subyacentes son lo suficientemente profundas como para aumentar sustancialmente la mortalidad.

2.2. Escalas pronósticas

Actualmente, existen una gran variedad de escalas pronósticas que permiten estandarizar y comparar datos clínicos. A partir de la puntuación obtenida de una escala

pronóstica, es posible predecir hasta cierto punto la evolución y el riesgo de mortalidad de un paciente [19]. Las escalas pronósticas son, por tanto, una herramienta predictiva de apoyo al diagnóstico muy importante en las Unidades de Cuidados Intensivos (UCI) de todo el mundo.

Como se comentó anteriormente, la escala que se utilizaba inicialmente para la identificación de la sepsis eran los criterios SIRS, pero actualmente su uso ya no está recomendado. En su lugar se emplea la escala SOFA.

Por otro lado, existen otras escalas más generales que pueden aplicarse para evaluar el riesgo de los pacientes. De entre ellas, destacan la Fisiológica Aguda y Evaluación Crónica de la Salud (*Acute Physiology and Chronic Health Evaluation,* APACHE), la Puntuación Simplificada Aguda Fisiológica (*Simplified Acute Physiology Score,* SAPS) y los Modelos de Probabilidad de Mortalidad (*Mortality Probability Models,* MPM) [19]. Estas escalas se describen brevemente en los siguientes apartados.

2.2.1. SIRS

Los criterios SIRS se emplean para describir la respuesta fisiopatológica a un ataque al cuerpo del paciente, como puede ser una infección, un traumatismo, quemaduras, una pancreatitis u otras lesiones [20]. Este concepto fue definido en 1991 tras el primer consenso mundial sobre la definición de la sepsis.

Se evalúan cuatro variables clínicas distintas para detectar la presencia de SIRS: temperatura, frecuencia cardiaca, frecuencia respiratoria y número de glóbulos blancos. Los umbrales definidos para cada una de ellas [21] son los siguientes:

- Temperatura < 36 °C o > 38°C.
- Frecuencia respiratoria > 90 latidos por minuto.
- Frecuencia respiratoria > 20 alientos por minuto.
- Recuento de glóbulos blancos < 4.000 células por mm^3 o > 12.000 células por mm^3.

Si se cumplen al menos dos de estos criterios, entonces se determina la presencia de SIRS.

Existen tres grandes inconvenientes con la definición SIRS [20]. En primer lugar, es una escala demasiado sensible, por lo que la mayoría de los pacientes de la UCI cumplen con sus criterios. En segundo lugar, no permite diferenciar una respuesta beneficiosa del

paciente de una respuesta patológica que pueda producir la disfunción orgánica. Por último, es muy difícil determinar el papel de la infección en la respuesta inflamatoria, así como establecer si esta respuesta se debe a factores no infecciosos. Por estos y otros motivos, los criterios SIRS están en desuso actualmente, y en su lugar se emplea la nueva escala SOFA.

2.2.2. SOFA

Como se ha comentado, el método más extendido actualmente para la evaluación de la severidad de la disfunción orgánica es la escala SOFA. Cuanto mayor es la escala SOFA, mayor es el riesgo de mortalidad. Por ejemplo, para una escala SOFA mayor o igual a 2 puntos, el riesgo de mortalidad para pacientes hospitalizados es superior al 10% [1]. En la Tabla 1 se muestran las diferentes variables clínicas que se evalúan a la hora de realizar el cálculo SOFA.

Tabla 1. Criterios de la escala SOFA. Fuente: [22]

	0	1	2	3	4
Respiración[a] PaO_2/FIO_2 (mm Hg) o SaO_2/FIO_2	>400	<400 221–301	<300 142–220	<200 67–141	<100 <67
Coagulación Plaquetas $10^3/mm^3$	>150	<150	<100	<50	<20
Hígado Bilirubina (mg/dL)	<1,2	1,2–1,9	2,0–5,9	6,0–11,9	>12,0
Cardiovascular[b] Tensión arterial	PAM ≥70 mmHg	PAM <70mm Hg	Dopamina a <5 o dobutamina a cualquier dosis	Dopamina a dosis de 5,1-15 o Epinefrina a ≤ 0,1 o Norepinefrina a ≤ 0,1	Dopamina a dosis de >15 o Epinefrina > 0,1 o Norepinefrina a > 0,1
Sistema Nervioso Central Escala de Glasgow	15	13–14	10–12	6–9	<6
Renal Creatinina (mg/dL) o flujo urinario (mL/d)	<1,2	1,2–1,9	2,0–3,4	3,5–4,9 <500	>5,0 <200

PaO_2: presión arterial de oxígeno; FIO_2: fracción de oxígeno inspirado; SaO_2, Saturación arterial de oxígeno periférico; PAM, presión arterial media; [a]PaO_2/FIO_2 es relación utilizada preferentemente, pero si no esta disponible usaremos la SaO_2/FIO_2; [b]Medicamentos vasoactivos administrados durante al menos 1 hora (dopamina y norepinefrina como ug/kg/min) para mantener la PAM por encima de 65 mmHg.

El cálculo de la puntuación SOFA requiere medidas de laboratorio, lo que dificulta su uso en entornos no hospitalarios. Debido a esto, se ha propuesto el uso de una nueva escala denominada qSOFA (*quick* SOFA) como alternativa a SOFA [1]. La escala qSOFA emplea únicamente criterios clínicos que se pueden medir de forma rápida y sencilla, por lo que puede ser usada incluso fuera de los hospitales. Los criterios empleados por qSOFA son tres:

- Frecuencia respiratoria ≥ 22/min
- presión arterial sistólica ≤ 100 mm Hg
- Puntuación en la escala Glasgow ≤ 15.

Diversos estudios [10] [11] demuestran que qSOFA tiene menor sensibilidad a la hora de identificar pacientes sépticos frente a otros criterios como SOFA o SIRS, por lo que no es recomendable emplear qSOFA como única herramienta para evaluar a los pacientes, sobre todo en entornos hospitalarios donde existe la posibilidad de realizar medidas más avanzadas. No obstante, la aparición de al menos dos de estos síntomas puede identificar rápidamente a pacientes en riesgo, y puede dar lugar a que los médicos tomen medidas adicionales para monitorizar al paciente más atentamente.

2.2.3. APACHE

La escala APACHE surgió en 1981 como una herramienta predictiva para determinar la gravedad de una enfermedad. Posteriormente, en 1985 se publicó APACHE II, una versión mejorada de la versión inicial. La escala APACHE II se calcula a partir de 12 parámetros fisiológicos medidos durante las primeras 24 horas de ingreso al hospital, además de tomar en cuenta otros factores como la edad y posibles enfermedades crónicas del paciente [25]. Estos criterios se muestran en la Tabla 2.

Tabla 2 Criterios de la escala APACHE II. Fuente: [26]

Variables Fisiológicas	**Rango elevado**					**Rango Bajo**				
	+4	**+3**	**+2**	**+1**	**0**	**+1**	**+2**	**+3**	**+4**	**Puntos**
Temperatura - rectal (°C)	≥41°	39 a 40.9°		38.5 a 38.9°	36 a 38.4°	34 a 35.9°	32 a 33.9°	30 a 31.9°	≤29.9°	
Presión arterial media (mmHg)	≥160	130 a 159	110 a 129		70 a 109		50 a 69		≤49	
Frecuencia cardiaca (respuesta ventricular)	≥180	140 a 179	110 a 139		70 a 109		55 a 69	40 a 54	≤39	
Frecuencia respiratoria (no ventilado o ventilado)	≥50	35 a 49		25 a 34	12 a 24	10 a 11	6 a 9		≤5	
Oxigenación : Elegir a o b a. Si FiO2 ³0,5 anotar P A-aO2 b. Si FiO2 < 0,5 anotar PaO2	≥500	350 a 499	200 a 349		<200 PO2>70	 PO2 61 a 70		 PO2 55 a 60	 PO2<55	
pH arterial (Preferido) HCO3 sérico (venoso mEq/l)	≥7.7 ≥52	7.6 a 7.69 41 a 51.9		7.5 a 7.59 32 a 40.9	7.33 a 7.49 22 a 31.9		7.25 a 7.32 18 a 21.9	7.15 a 7.24 15 a 17.9	<7.15 <15	
Sodio Sérico (mEq/l)	≥180	160 a 179	155 a 159	150 a 154	130 a 149		120 a 129	111 a 119	≤110	
Potasio Sérico (mEq/l)	≥7	6 a 6.9		5.5 a 5.9	3.5 a 5.4	3 a 3.4	2.5 a 2.9		<2.5	
Creatinina sérica (mg/dl) Doble puntuación en caso de fallo renal agudo	≥3.5	2 a 3.4	1.5 a 1.9		0.6 a 1.4		<0.6			
Hematocrito (%)	≥60		50 a 59.9	46 a 49.9	30 a 45.9		20 a 29.9		<20	
Leucocitos (Total/mm3 en miles)	≥40		20 a 39.9	15 a 19.9	3 a 14.9		1 a 2.9		<1	
Escala de Glasgow Puntuación=15-Glasgow actual										
A. APS (Acute Physiology Score) Total: Suma de las 12 variables individuales										
B. Puntuación por edad (≤44 = 0 punto; 45-54 = 2 puntos; 55-64 = 3 puntos; 65-74 = 5 puntos; ≥75 = 6 puntos)										
C. Puntuación por enfermedad crónica (ver más abajo)										
Puntuación total APACHE II (Suma de A+B+C)										

Existe una fuerte correlación entre la puntuación calculada a partir de los criterios APACHE II y la tasa de mortalidad apreciada en los pacientes [27]. En la Tabla 3 se

muestra la tasa de mortalidad esperada en pacientes quirúrgicos y no quirúrgicos en base a la puntuación APACHE II calculada. Se puede observar que, cuanto mayor es la puntuación, mayor es el riesgo.

Tabla 3. Puntuación APACHE II y mortalidad esperada. Fuerte: [27].

	Mortalidad esperada (%)	
Puntuación	**Pacientes quirúrgicos**	**Pacientes no quirúrgicos**
0-4	2	4
5-9	4	8
10-14	8	12
15-19	12	25
20-24	29	40
25-29	35	50
30-34	70	70
> 34	88	80

Existe una versión mejorada de APACHE II denominada APACHE III. Las principales diferencias entre ambas son el valor de la puntuación total, el número de variables fisiológicas analizadas (APACHE III tiene en cuenta 17 parámetros) y la evaluación de enfermedades crónicas. En general, APACHE III no ofrece una mejora sustancial frente a APACHE II [28], pero hay que destacar que ambas vertientes han probado ser altamente competitivas a la hora de predecir la mortalidad en pacientes con sepsis [29].

2.2.4. SAPS

SAPS es una versión simplificada de APACHE que permite estimar la gravedad de un paciente mediante la valoración de datos clínicos simples y fácilmente medibles [18]. Al igual que APACHE, SAPS cuenta con diferentes revisiones. Hoy en día, la versión más usada es SAPS II.

SAPS evalúa los peores valores de 17 variables clínicas medidos durante las primeras 24 horas desde el ingreso a la UCI [19]. Estas variables pueden ser de dos tipos básicos: variables: binarias o continuas. En la Tabla 4 se muestran, de forma simplificada, los criterios SAPS II y la puntuación asociada a cada variable clínica según el valor que toma.

Tabla 4. Criterios de SAPS II. Fuente: [30]

Variables	Puntuación													
	26	13	12	11	9	7	6	5	4	3	2	0	1	10
HR (latidos/min)			< 40			≥ 160			120-159		40-69	70-119		
SBP (mmHg)		< 70						70 - 99			≥ 200	100-119		
Temperatura										≥ 39		< 39		
Respiración (PaO_2/FiO_2)				< 100	100-99	≥ 200								
Orina (L/día)				< 0,5					0,5 - 0,99			≥ 1		
Urea (g/L)							0,6 - 1,7					< 0,6		> 1,8
TLC			< 1								≥ 20	1 – 19,9		
Potasio										< 3	≥ 5	3-4,9		
Sodio								< 125				125-144	≥ 145	
Bicarbonato							< 15			15-19		> 20		
Bilirrubina (mg/dl)					≥ 60				40-59,9			< 40		
GCS	< 6	6-8				9-10		11-13				15-15		

Edad	Puntuación	Enfermedad crónica	Puntuación	Tipo de admisión	Puntuación
< 40	0	Cáncer metastático	9	Cirugía programada	0
40 - 59	7	Neoplastias hematológicas	10	Médica	6
60 – 69	12	SIDA	17	Cirugía de emergencia	8
70 -74	15				
75 – 79	16				
> 80	18				

HR: frecuencia cardiaca; SBP: presión sanguínea sistólica; TLC: Recuento total de leucocitos; GCS: Puntuación de coma Glasgow.

Al igual que las escalas anteriores, una mayor puntuación SAPS está relacionada con una mayor tasa de mortalidad. En la 5 se muestra una estimación de la mortalidad esperada a partir de la puntuación SAPS obtenida.

Tabla 5. Relación entre puntuación SAPS y mortalidad esperada. Fuente: [18]

Puntuación	Mortalidad esperada (%)
5-6	10,7±4,1
7-8	13,3± 3,9
9-10	19,4± 7,8
11-12	24,5± 4,1

13-14	30,0± 5,5
15-16	32,1± 5,1
17-18	44,2± 7,6
19-20	50,0± 9,4
≥21	81,1± 5,4

2.2.5. MPM

MPM es una escala pronóstica que utiliza variables clínicas simples medidas al momento del ingreso y tras 24 horas en la UCI. Tiene en cuenta, además, la edad del paciente y su estado de salud previo a la hospitalización [19].

En la Tabla 6 se muestran los datos recogidos en el momento de la entrada del paciente a la UCI según MPM II. Por otro lado, en la Tabla 7 se enumeran las variables que se analizan una vez han transcurrido 24 horas. Como se puede observar, se analizan nuevamente las variables de entrada para evaluar su evolución tras el tratamiento.

Tabla 6. Variables MPM II medidas al momento del ingreso. Fuente: [18]

Variables medidas al momento del ingreso	
Edad	
Alteración fisiológica aguda: − Coma o estupor. − Frecuencia cardiaca ≥ 150 ppm. − Tensión arterial sistólica ≤ 90 mmHg. − Ventilación mecánica. − Fracaso Renal Agudo.	− Arritmias cardiacas graves. − Accidente cerebrovascular. − Sangrado gastrointestinal − Efecto masa craneal. − Reanimación cardiopulmonar previa al ingreso.
Estado crónico de salud: − Insuficiencia renal crónica. − Cirrosis. − Neoplasia metastásica.	
Tipo de paciente: − Paciente médico. − Paciente quirúrgico urgente.	

Tabla 7. Variables MPM II medidas 24 horas tras el ingreso. Fuente: [18]

Variables 24 horas después del ingreso	
Edad	
Variables medidas al ingreso − Insuficiencia renal crónica. − Cirrosis. − Neoplasia metastásica. − Tipo de paciente.	
Parámetros evaluados tras tratamiento: − Coma o estupor profundo a las 24 horas. − Creatinina > 2 mg/dl. − Infección confirmada. − Ventilación mecánica a las 24 horas del ingreso. − PO2 < 60 mmHg. − Tiempo de Protrombina.	

MPM II evalúa la presencia de cada una las variables bajo estudio y les asigna una puntuación en función de su peso estadístico. Esta metodología permite hacer una estimación directa de la probabilidad de mortalidad. Si se realiza una medida diaria adicional cada 24 horas, es posible determinar si el paciente está respondiendo de forma adecuada al tratamiento aplicado. Un paciente que mantiene coeficientes estables a pesar de estar bajo tratamiento incrementa la probabilidad de mortalidad significativamente [18].

2.3. Gestión de la sepsis

La detección y gestión temprana de la sepsis es un factor clave a la hora de prevenir la mortalidad de los pacientes, ya que facilita el tratamiento temprano de la enfermedad. Por tanto, las primeras horas de la enfermedad son las más importantes. Debido a la variedad de síntomas asociados a la sepsis, es habitual que se realice el diagnóstico antes de que los resultados de los análisis de sangre estén disponibles. Para diagnosticar la sepsis, los médicos deben obtener el historial clínico del paciente y examinar los resultados de laboratorio en busca de síntomas de infección. Generalmente, el sistema respiratorio es la ubicación más común para la aparición de la enfermedad, y la fiebre es el síntoma que se da con más frecuencia. Si se sospecha la presencia de sepsis, se procede a verificar si se cumplen o no los criterios requeridos por la escala pronóstica en uso y se comienza la monitorización continua del paciente. Esto conlleva la evaluación de las vías respiratorias, la respiración y la circulación. Es necesario obtener estudios de laboratorio e imagen para identificar el origen de la infección [31].

Posteriormente, se inicia el tratamiento. En este aspecto, destaca la terapia temprana dirigida por objetivos, un protocolo diseñado para implementarse durante las seis primeras horas desde la identificación de un posible caso de sepsis o *shock* séptico. Este protocolo requiere, en primer lugar, la estabilización respiratoria del paciente. Para pacientes de alto riesgo, se debe colocar un catéter venoso central o bien iniciar la ventilación mecánica. Una vez que se ha realizado la estabilización respiratoria inicial, el tratamiento consiste en reanimación con fluidos, terapia de agentes vasopresores, identificación y control de la infección, administración rápida de antibióticos, y eliminación o drenaje de la fuente de infección [31]. La terapia temprana dirigida por objetivos puede reducir de forma significativa la mortalidad de pacientes sépticos hospitalizados en la UCI (30%) en comparación con un tratamiento estándar (46%) [32].

2.4. Modelos de predicción basados en Machine Learning aplicados a la detección de sepsis

La detección temprana de la sepsis constituye uno de los principales factores para conseguir reducir la mortalidad y secuelas médicas asociadas a esta afección. El principal reto al que se enfrentan los médicos a la hora de realizar un diagnóstico precoz es distinguir la sepsis de otros problemas clínicos con síntomas similares. Ésta es también la causa por la que actualmente no existe un método universalmente aceptado para el diagnóstico de la sepsis. En los últimos años, junto con el auge de la tecnología y el *Big Data*, se han creado registros médicos digitales que contienen una gran cantidad de datos clínicos de alta calidad y resolución. Los modelos basados en Machine Learning son capaces de usar gran cantidad de datos complejos y aprender patrones útiles que permitan realizar predicciones sobre la evolución de los pacientes y la probabilidad de que desarrollen sepsis.

Recientemente, una revisión sistemática [33] ha concluido que los modelos de Machine Learning son capaces de predecir la aparición de sepsis con alta sensibilidad y especificidad. Las metodologías de aprendizaje profundo tienen un rendimiento especialmente bueno a la hora de realizar la detección de la enfermedad. Una reciente comparación entre un nuevo modelo de detección de las fases iniciales de la sepsis basado en aprendizaje profundo y un modelo de regresión con extracción de características temporales convencionales [34], concretamente el modelo InSight [35], concluyó que el uso de modelos de aprendizaje profundo proporciona una mejora en la capacidad de

análisis y predicción de la sepsis sin necesidad de hacer una extracción de características de los datos en bruto. No obstante, el estudio también reconoce que estos modelos tienen limitaciones en cuanto a la fijación de las variables a usar, tener una cantidad de datos suficiente y la presentación de la información en un formato que sea fácilmente interpretable para los médicos.

En línea con lo anterior, en 2019 se lanzó el "The PhysioNet/Computing in Cardiology Challenge 2019" [14]. El objetivo de dicho reto era la detección temprana de sepsis seis horas antes del diagnóstico usando datos fisiológicos. La evaluación de los participantes se llevó a cabo mediante las métricas estándar (exactitud, sensibilidad, especificidad, valor F1, etc.), así como usando una nueva métrica específica del reto, denominada función de utilidad. La función de utilidad valora la utilidad real del algoritmo para los médicos en función de con cuánta antelación ha sido capaz de predecir la sepsis. El valor máximo normalizado de la utilidad es de 1,0 para los pacientes que desarrollan sepsis eventualmente. Para pacientes no sépticos, el valor máximo es 0.

En la Tabla 8 se muestran los ocho primeros clasificados del reto y el tipo de modelo basado en Machine Learning empleado por cada uno. La lista completa de resultados está disponible en [14]. Es necesario precisar que la clasificación se hizo en términos de "utilidad", pues refleja muy bien qué espera un profesional sanitario de cara a la detección temprana. En consecuencia, es una de las métricas principales de este proyecto, así como las estándar.

Tabla 8. Ocho primeros clasificados del reto de PhysioNet.

Clasificación	Modelo empleado	Utilidad
1	Modelo de regresión basado en la firma de características [36]	0,360
2	Redes neuronales	0,345
3	Conjunto de modelos XGBoost [37]	0,339
4	Modelo Time-Phased [38]	0,337
5	Modelo basado en XGBoost [39]	0,337
6	Árboles de decisión basados en refuerzo [40]	0,332
7	Modelo basado en XGBoost	0,331
8	Arquitectura neuronal [41]	0,328

Capítulo 3. Marco metodológico

3.1. Machine Learning

La Inteligencia Artificial puede entenderse como una rama de la ciencia con la que se busca imitar la capacidad humana de aprender y hacer razonamiento lógico a partir del aprendizaje. Dentro de la Inteligencia Artificial se encuentra el Machine Learning, un campo que pretende dotar a los ordenadores de la capacidad de aprender sin necesidad de haber sido programados previamente [42].

Las técnicas de Machine Learning pueden emplearse para tareas de descripción, predicción e inferencia causal [43]. La descripción consiste en usar datos para proporcionar un resumen cuantitativo basado en variables específicas. Buscar la proporción de pacientes con una característica concreta en un conjunto de datos es un ejemplo de descripción. Por otro lado, la predicción se basa en la regresión o clasificación de un resultado de interés, como el pronóstico del riesgo o la evolución de un paciente en función de unas variables de entrada. Por último, la inferencia causal implica estimar efectos sobre el conjunto de datos. Esto se realiza comparando los resultados presentados con una determinada condición con los resultados contrafactuales si no se hubiera dado dicha condición. Un ejemplo de inferencia causal podría ser estimar la efectividad de un tratamiento médico comparando la tasa de recuperación obtenida con y sin su aplicación.

Debido a la gran potencia que ofrece Machine Learning, en la actualidad su uso se ha extendido a una gran variedad de ámbitos. En 2020, una encuesta llevada a cabo por Deloitte [44] reveló que el 67% de las empresas ya usaban técnicas de Machine Learning, y el 97% planeaban usarlas el siguiente año. Según una reciente revisión [45], las principales áreas de aplicación de Machine Learning son visión por ordenador, predicción, análisis semántico, procesado del lenguaje natural y obtención de información. Dentro de estas áreas hay, a su vez, áreas más específicas en las que se emplea Machine Learning para realizar tareas concretas como pueden ser el reconocimiento y detección de objetos o la clasificación. En la Figura 1 se muestra un esquema completo de estas áreas de aplicación.

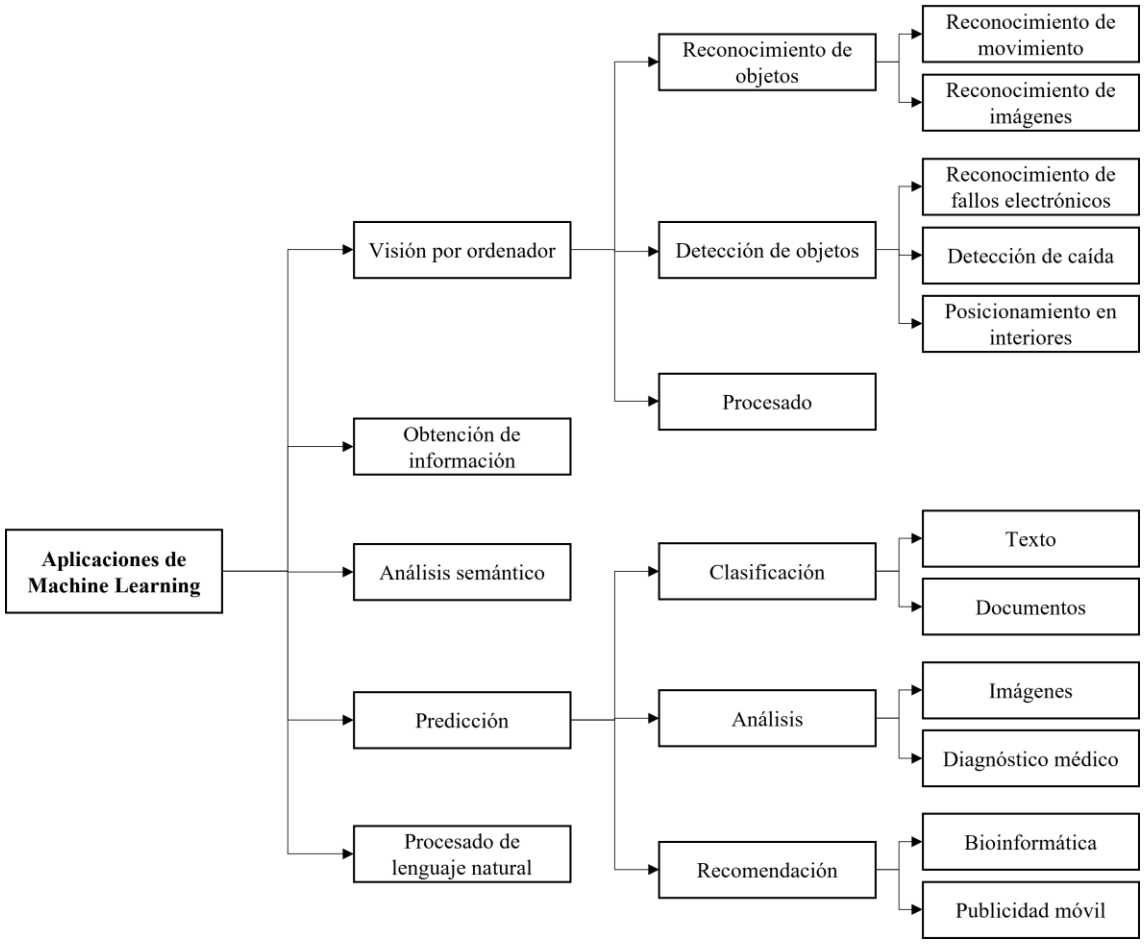

Figura 1. Aplicaciones de Machine Learning. Fuente: [45].

Dentro de las disciplinas de Machine Learning existentes se encuentra el Deep Learning. Éste permite que modelos computacionales formados por múltiples capas de procesamiento puedan aprender representaciones de datos con múltiples niveles de abstracción [46]. Por otro lado, las aplicaciones de Deep Learning están presentes en nuevas áreas como procesado de imágenes médicas, robótica, coches autónomos y Big Data en general. Una de las aplicaciones recientes de esta modalidad de inteligencia artificial es la gestión de relaciones con los clientes, que se basa realizar un análisis del historial de un usuario que se usa para mejorar la relación comercial [45]. Asimismo, el Deep Learning es capaz de mejorar el rendimiento de las aplicaciones de Machine Learning anteriormente mencionadas, y ha demostrado ser más eficiente en diferentes tareas, como la reconstrucción de circuitos cerebrales [47] y la predicción de los efectos de mutaciones del ADN no codificante sobre la expresión de los genes y enfermedades [48], entre otros.

Actualmente, la popularidad de Deep Learning está en auge gracias al gran número de aplicaciones que tiene y los buenos resultados que alcanza. No obstante, esto ha llevado a un uso excesivo de estas tecnologías sin considerar si son realmente necesarias para el problema que se afronta. De acuerdo con varios autores [49][50], lo recomendable es comenzar empleando modelos de inteligencia artificial convencionales o no profundos antes de considerar el uso del aprendizaje profundo, ya que generalmente son más rápidos, simples, menos costosos computacionalmente y de probada bondad en diversas aplicaciones. Esta visión es la empleada para este TFM. En coherencia con esta visión, se ha decidido comenzar por explorar el uso del clasificador de Bayes como una de las técnicas de Machine Learning de primera elección. Constituye un "clasificador base". Este clasificador se explica en la sección 3.3, y otros "base" en la sección 3.7. Esta visión se completa con lo expresado en el apartado 3.5 donde se verá cómo combinar diversos clasificadores base.

3.2. Modelos de aprendizaje

Los algoritmos de Machine Learning deben pasar por un proceso de aprendizaje. Actualmente, existen diferentes modalidades de aprendizaje en Machine Learning. Los tres paradigmas básicos son aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo. En los siguientes apartados se describen brevemente las principales características de cada uno de estos modelos.

3.2.1. Aprendizaje supervisado

El aprendizaje supervisado consiste en entrenar al algoritmo con datos etiquetados de entrada y salida en los que se conoce correctamente su relación. A partir de estos ejemplos, la máquina aprenderá diferentes patrones y construirá una función de clasificación propia. Algunas técnicas de Machine Learning más entrenadas mediante aprendizaje supervisado son el árbol de decisión, las Máquinas de Vectores de Soporte (*Support Vector Machines,* SVM) y Naïve Bayes [51].

En la Figura 2 se muestra el diagrama de flujo general para un modelo de aprendizaje supervisado. Los datos se dividen en conjuntos de entrenamiento y prueba. Estos conjuntos de datos se usan para entrenar y evaluar el modelo, respectivamente. Una vez que se ha validado el modelo, se puede aplicar a nuevos conjuntos de datos para realizar clasificaciones o predicciones. Normalmente, los modelos están sometidos a un proceso

de mejora continuo, en los que se afina su funcionamiento en base a los resultados obtenidos.

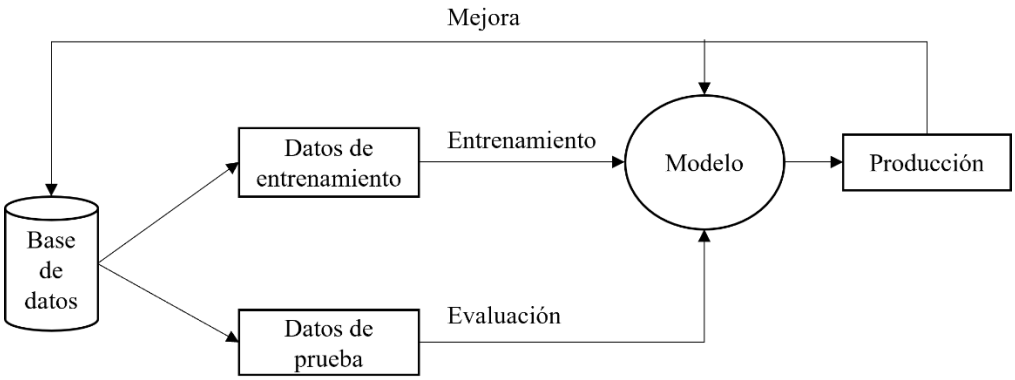

Figura 2. Diagrama básico del aprendizaje supervisado. Fuente: [51].

El aprendizaje supervisado requiere de asistencia humana, ya que debe haber un supervisor encargado de dividir los datos en los conjuntos de entrenamiento y prueba, de y etiquetarlos correctamente, además de verificar la salida obtenida.

3.2.2. Aprendizaje no supervisado

La modalidad de aprendizaje no supervisado se basa en que no existe información de entrada sobre el resultado correcto o esperado, es decir, los datos no están etiquetados. Como su nombre indica, no existe intervención humana. El algoritmo tiene la tarea de reconocer y presentar la información relevante extraída de los datos. Cuando se le inyectan nuevos datos, el algoritmo hace uso de las características aprendidas previamente para reconocer la clase a la que pertenecen los datos [51].

El aprendizaje no supervisado emplea dos técnicas básicas:

- ***Clustering:*** es un proceso que consiste en dividir un conjunto de datos en diferentes subclases o grupos. Cada grupo está formado por una colección de objetos similares que son tratados colectivamente. Un algoritmo de *clustering* es bueno cuando la similitud de los datos agrupados es alta y los grupos están bien diferenciados entre sí. Normalmente, la similitud entre clases se mide por proximidad [52].
- **Reducción de la dimensionalidad:** son procedimientos que generan subespacios de datos de menor dimensión derivados del conjunto original. En estos subespacios está contenida la mayor parte de la información útil, pero con un número reducido de variables, lo que facilita la comprensión de los datos y la

eliminación de redundancias y permite mejorar la clasificación y visualización de los datos con un menor coste computacional [53].

Entre las tecnologías que se entrenan mediante aprendizaje no supervisado destacan Análisis del Componente Principal (*Principal Component Analysis*, PCA) y K-Means [51].

3.2.3. Aprendizaje por refuerzo

En el aprendizaje por refuerzo, el algoritmo interactúa con el entorno a través de diferentes acciones. Estas acciones afectan al entorno, de forma que la máquina recibe castigos o recompensas. El objetivo es que el algoritmo aprenda a trabajar de forma que maximice el número de recompensas recibidas [54].

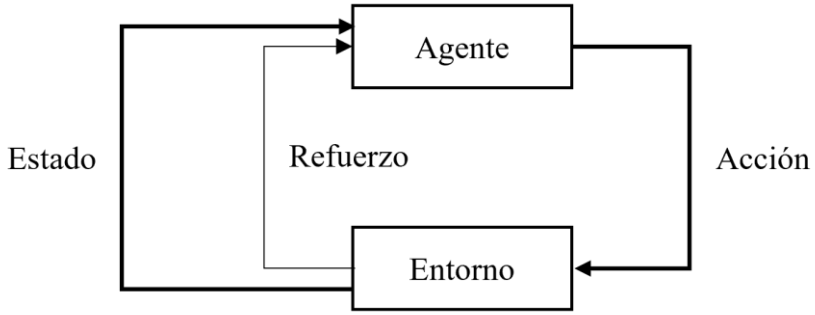

Figura 3. Diagrama básico del aprendizaje por refuerzo. Fuente: [51].

Un modelo de aprendizaje por refuerzo estándar consta de un agente conectado al entorno a través de acciones, como muestra la Figura 3. En cada iteración, el agente recibe el estado actual del entorno y elige realizar una acción. Esta acción modifica el entorno, y el valor de la transición se comunica al agente en forma de una señal de refuerzo escalar. Las acciones elegidas por el agente deben intentar maximizar la suma del conjunto de señales de refuerzo obtenidas. El proceso de aprendizaje se realiza mediante prueba y error [55].

3.3. Teorema de Bayes

El teorema de Bayes fue planteado por primera vez en 1763 por el matemático inglés Thomas Bayes, y se basa en el uso de probabilidades condicionadas. Se puede entender el teorema de Bayes como un proceso de aprendizaje, donde la transición de la probabilidad *a priori* de la hipótesis a la probabilidad *a posteriori* refleja lo aprendido

sobre la validez de la hipótesis tomando en cuenta la información extraída de los datos [56]. De forma general, este teorema se expresa con la siguiente ecuación:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (1)$$

Donde:

- $P(Y)$ y $P(X)$son las probabilidades *a priori* de la hipótesis Y y los datos X, respectivamente. Reflejan los conocimientos que se tienen previamente, puede ser incluso antes de tener en cuenta los datos.
- $P(Y|X)$ es la probabilidad de que la hipótesis Y sea cierta si se tienen los datos X.
- $P(X|Y)$ es la probabilidad de que, siendo la hipótesis Y cierta, se hayan dado los datos X. También se conoce como hipótesis *a posteriori*. Normalmente, esta probabilidad es conocida, y se puede incluir en el proceso de aprendizaje, ya que representa el estado esperado de los datos cuando la hipótesis planteada es verdadera [57].

En el ámbito que atañe a este proyecto, desde una óptica científico-tecnológica, es la obtención de $P(Y)$, $P(X)$ y $P(X|Y)$ lo que más se trabaja, a la vez que desde una óptica científico-médica lo que se espera es la probabilidad de que la hipótesis Y sea cierta a partir de los datos X. Por decirlo de otra forma, de cara a la aplicación práctica, los ingenieros se encargan del diseño de los tests (por ejemplo, hipótesis de si hay o no sepsis) y los médicos de su uso. Lógicamente, los aspectos científicos son interdisciplinares. Este aspecto se ha tenido en cuenta a la hora de definir el reto de PhysioNet.

El concepto de las probabilidades *a priori* es el más complejo dentro de la teoría Bayesiana. Existen dos vertientes principales para el cálculo de estas probabilidades: frecuentista y subjetivista [58]. La primera consiste en obtener la probabilidad de un evento específico a partir de su frecuencia relativa dentro del número de eventos observados. En esta situación, y en ausencia de más información, la probabilidad frecuentista asociada con la observación de un caso en la población de interés es la prevalencia de los datos. Por otro lado, el enfoque subjetivista solo requiere que alguien (por ejemplo, un médico) pueda establecer cómo de probable es una hipótesis. Cualquiera de los dos métodos puede ser usado para calcular las probabilidades *a priori* de los datos, siempre y cuando exista una base de conocimiento previo (por ejemplo, médico) empírico o teórico lo suficientemente fiable como para respaldarlas.

El objetivo de un clasificador Bayesiano es obtener la clase con la probabilidad más alta, es decir, la hipótesis *Maximum a Posteriori* (MAP):

$$y_{MAP} = \arg máx_{y \in Y}(P(X_n|y)P(y)) \quad (2)$$

donde $X_n$es el vector de variables o datos. La estimación directa de la probabilidad más alta a partir de la ecuación (2) se vuelve demasiado compleja cuando existe un elevado número de características X_n. Debido a esto, es común usar diferentes aproximaciones que faciliten el cálculo, como asumir que las características de una clase son independientes entre sí. Este principio es usado por el clasificador de Naïve Bayes.

3.3.1. Clasificador Naïve Bayes

Un clasificador de Naïve Bayes asume la independencia entre las características de una clase. Esto quiere decir que cada característica tiene únicamente la clase como raíz (Figura 4) [59]. Por ejemplo, se puede identificar una fruta como manzana si es roja, redonda y tiene alrededor de ocho centímetros de diámetro. Aunque en realidad estas características están relacionadas entre sí, en una aproximación basada en el clasificador de Naïve Bayes se considerará que todas estas propiedades contribuyen de manera independiente a la probabilidad de que la fruta sea una manzana [60]. Por tanto, la ecuación (2) se puede reescribir como la siguiente expresión:

$$P(Y|X) = P(Y)\prod_{n=1}^{N} P(X_n|Y) \quad (3)$$

Gracias a su gran eficiencia y fiabilidad, los clasificadores de Naïve Bayes se emplean actualmente en diversas aplicaciones, como diagnóstico médico, sistemas de recomendación, filtros de correo electrónico, optimización de páginas web y detección de fraudes, entre otros [59]. En este punto hemos de comentar que se usará un clasificador Naïve, pero sí que se aprovecharán sus potencialidades, tal como se explicará en el apartado 3.4.

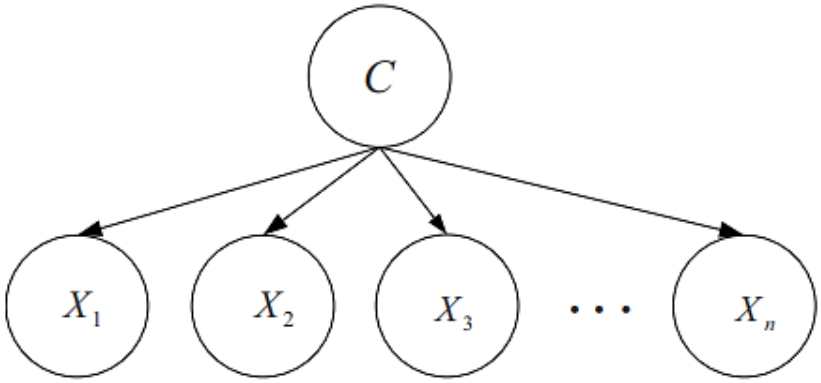

Figura 4. Independencia de características en Naïve Bayes. Fuente: [59].

3.3.2. Evaluación de test diagnóstico

Generalmente, se usan dos tipos de test en la práctica clínica: cribado y diagnóstico. Los test de cribado son aquellos que se usan en los pacientes sanos o asintomáticos para determinar si es necesaria una intervención o identificar una enfermedad precozmente. Los test de cribado son, por ejemplo, aquellos que se realizan durante un reconocimiento médico rutinario. Por otro lado, los test de diagnóstico sirven de ayuda para tomar decisiones clínicas, por lo que se usan en pacientes que ya presentan síntomas, o bien después de un test de cribado positivo para establecer un diagnóstico definitivo. La correcta interpretación de los resultados puede convertirse en un reto, ya que depende de la habilidad discriminante del test [61], es decir, de la habilidad de distinguir entre las dos categorías de interés, salud o enfermedad en este caso.

Para evaluar la fiabilidad de un test de diagnóstico o cribado, se pueden emplear métricas estándar como la sensibilidad, la especificidad o la exactitud. Sin embargo, aunque dichas métricas son útiles y permiten saber la calidad del test, lo que en muchas ocasiones los profesionales médicos realmente quieren saber es la probabilidad de tener la enfermedad teniendo un resultado concreto en el test, es decir, la capacidad predictiva del test. En general, las cuestiones principales que hay que plantearse son las siguientes:

1. Sabiendo que el paciente tiene la enfermedad, ¿cuál es la probabilidad de tener un test positivo, es decir, $P\{T|D\}$?
2. Teniendo un test positivo, ¿cuál es la probabilidad de que el paciente tenga la enfermedad, es decir, $P\{D|T\}$?
3. Sabiendo que el paciente no tiene la enfermedad, cuál es la probabilidad de tener un test negativo, es decir, $P\{\overline{T}|\overline{D}\}$?

4. Teniendo un test negativo, ¿cuál es la probabilidad de que el paciente no tenga la enfermedad, es decir, $P\{\bar{D}|\bar{T}\}$?

En la Tabla 9 se muestran, de forma resumida, las medidas estadísticas para la evaluación de un test diagnóstico en forma de probabilidad condicional. Hay que destacar que la Fracción de Verdaderos Positivos (FVP) se corresponde con la sensibilidad, mientras que la Fracción de Verdaderos Negativos (FVN) se corresponde con la especificidad.

Tabla 9. Medidas estadísticas para evaluar un test de diagnóstico (I). Fuente: [62]

Resultado del test	**Estado de la enfermedad**	
	Enfermo (D)	**Sin enfermedad ($\bar{D}$)**
Positivo (T)	Fracción de Verdaderos Positivos $P\{T\|D\}$	Fracción de Falsos Positivos $P\{T\|\bar{D}\}$
Negativo ($\bar{T}$)	Fracción de Falsos Negativos $P\{\bar{T}\|D\}$	Fracción de Verdaderos Negativos $P\{\bar{T}\|\bar{D}\}$

Como se ha comentado anteriormente, lo que el personal médico desea conocer es cómo de buena es la predicción hecha por el test. En este aspecto, el "valor predictivo" [61] del test es fundamental, ya que permite definir cuánto de segura es la predicción. En la Tabla 10 se muestran los principales valores predictivos. El Valor Predictivo Positivo (VPP) es la probabilidad de que una persona con un test positivo tenga realmente la enfermedad. El Valor Predictivo Negativo (VPN), por su parte, es la probabilidad de que una persona que haya dado negativo en el test no tenga la enfermedad.

Tabla 10. Medidas estadísticas para evaluar un test de diagnóstico (II). Fuente: [62]

Resultado del test	**Estado de la enfermedad**	
	Enfermo (D)	**Sin enfermedad ($\bar{D}$)**
Positivo (T)	Valor Predictivo Positivo $P\{D\|T\}$	$P\{\bar{D}\|T\}$
Negativo ($\bar{T}$)	$P\{D\|\bar{T}\}$	Valor predictivo negativo $P\{\bar{D}\|\bar{T}\}$

En principio, no es posible obtener los valores predictivos del test a partir de los resultados de la clasificación como ocurre con la sensibilidad o la especificidad. Esto es debido a que VPP y VPN no son características estáticas del test, sino que dependen de la prevalencia de la enfermedad [61]. No obstante, si se conoce dicha prevalencia, $P\{D\}$, se pueden calcular los valores predictivos a partir de la sensibilidad y la especificidad usando el teorema de Bayes [62]. Así, para calcular VPP, se usa la siguiente expresión:

$$P\{D|T\} = \frac{P\{T|D\}P\{D\}}{P\{T|D\}P\{D\} + P\{T|\overline{D}\}P\{\overline{D}\}} \quad (4)$$

donde $P\{T|\overline{D}\}$ es 1 – especificidad, y $P\{\overline{D}\}$ es 1 - prevalencia. Por otro lado, VPN se calcula mediante la siguiente igualdad:

$$P\{\overline{D}|\overline{T}\} = \frac{P\{\overline{T}|\overline{D}\}P\{\overline{D}\}}{P\{\overline{T}|\overline{D}\}P\{\overline{D}\} + P\{\overline{T}|D\}P\{D\}} \quad (5)$$

donde $P\{\overline{T}|D\}$ es 1 – sensibilidad.

3.4. Validación cruzada

Uno de los principales retos a los que se enfrentan los algoritmos entrenados por aprendizaje supervisado es el *overfitting*, es decir, que el modelo esté perfectamente adaptado al conjunto de datos de entrenamiento y sea incapaz de generalizar cuando se le insertan datos nuevos. Las causas de este fenómeno pueden clasificarse en tres clases generales [63]:

- **Aprendizaje del ruido en los datos de entrenamiento**. Se da cuando el conjunto de datos de entrenamiento es muy pequeño o tiene pocos datos útiles. Estas circunstancias aumentan la probabilidad de que el modelo aprenda ruido y lo use posteriormente como base para la predicción.
- **Complejidad de las hipótesis**. Si hay demasiadas entradas (hipótesis), el modelo se vuelve más preciso, pero conseguir un clasificador consiste se hace más difícil, esto es, necesitamos muchos más datos. Una de las consecuencias es que el modelo será muy diferente para distintos conjuntos de datos.
- **Procedimientos de comparaciones múltiples**. Generalmente, un algoritmo de inteligencia artificial compara diferentes variables basándose en un sistema de evaluación y selecciona aquella que tenga la mayor puntuación. No obstante, durante este proceso probablemente se seleccionarán algunos parámetros que no mejoren la exactitud del modelo, o que incluso la empeoren.

Idealmente, se evaluaría el modelo usando datos nuevos que se hayan originado en la misma población que los datos de entrenamiento, pero en la práctica esto generalmente no es viable. En su lugar, para estimar la capacidad de generalización del modelo se suelen emplear métodos de remuestreo de datos, como la validación cruzada. El concepto básico de la validación cruzada es dividir los datos en distintos subconjuntos. Sobre este conjunto de datos disponibles para el aprendizaje se aplican métodos de remuestreo aleatorio para generar uno o más conjuntos de entrenamiento y validación. Existen diferentes métodos de validación cruzada [64]:

- **Submuestreo aleatorio *"single hold-out".*** Es una de las estrategias más simples. Consiste en seleccionar de forma aleatoria algunos casos para el subconjunto de validación, mientras que el resto de los casos se mantienen para el entrenamiento. Por lo general, se emplea entre el 10% y el 30% de los datos para el test y el resto para entrenamiento.
- **Submuestreo aleatorio *"k-fold".*** Consiste en repetir k veces el método de *single hold-out*, de forma que se generan k pares de subconjuntos. La función de aprendizaje se aplica a cada subconjunto de entrenamiento, y el modelo resultante se valida usando el subconjunto de test correspondiente. El rendimiento total se estima como la media de los modelos generados. En ocasiones también se da la desviación estándar.
- **Validación cruzada *"k-fold".*** El conjunto de datos se divide en k subconjuntos de aproximadamente el mismo tamaño. El modelo se entrena usando $k - 1$ subconjuntos, mientras que el subconjunto restante se usa para validación. Este proceso se repite hasta que todos los subconjuntos se hayan usado para validación.
- **Validación cruzada *"Leave-one-out".*** Es un caso especial de *k-fold* en el que se trata a cada caso individual como un subconjunto. Es decir, en un conjunto de datos con n pacientes habría $k = n$ subconjuntos. El primer subconjunto de validación contendría solamente el primer paciente, x_1, el segundo subconjunto contendría solamente el segundo paciente, x_2, y así sucesivamente hasta llegar al paciente x_n.

3.5. Estrategias de combinación

Cuando se trabaja con clasificadores, se puede dar que haya más de un clasificador atacando el mismo problema con los mismos datos. En una circunstancia así, sería deseable sacar lo mejor de cada uno, tanto de forma individual como conjunta. En este segundo caso, se puede plantear una estrategia de combinación de clasificadores como forma de obtener mejor clasificación que con cualquiera de los individuales. La visión anterior es aplicable a este caso, ya que se trabaja con diferentes variables clínicas que a su vez pueden dar lugar a clasificadores diseñados individualmente a partir de estas variables. Es el caso que ocupa este TFM. Si bien para este trabajo no es necesario hacer suposiciones acerca de la independencia entre las variables, sí que resulta interesante plantear la situación a partir de similitudes con el caso *Naïve* pues las analogías son casi directas y facilitan los planteamientos.

El clasificador Naïve Bayes trabaja con varias variables clínicas distintas, de cada una se obtendrá un "clasificador base". Para cada uno de ellos, se obtendrá una pareja de probabilidades asociadas a la clase positiva o negativa. Un método adecuado de combinación de parámetros podría permitir una mejora en la eficiencia del algoritmo, de forma que la clasificación sea más precisa que usando variables individuales. Para este trabajo, se han tenido en cuenta tres estrategias de combinación principales [65], singularizadas a un clasificador Bayesiano. Sea L el número total de variables clínicas usadas, c el número de clases (sepsis y no sepsis, en este caso), y $p_{i,j}$ $(i = 1, \dots, L; j = 1, \dots, c)$ las probabilidades asociadas al clasificador base *i-ésimo*. Entonces, se tienen estas estrategias:

- **Combinación lineal.** Es una combinación que consiste en calcular la media aritmética de las probabilidades $p_{i,j}$ de los diferentes parámetros usados en el clasificador, de manera que la clase asignada Y será aquella para la que la probabilidad obtenida sea mayor. La combinación lineal se puede definir con la siguiente expresión:

$$Y = \arg máx_{j=1}^{c} \frac{\sum_{i=1}^{L} p_{i,j}}{L} \qquad (6)$$

- **Combinación producto.** De forma análoga al caso anterior, este método consiste en calcular el producto de las probabilidades de las diferentes

variables clínicas usadas y asignar la clase para la cual el resultado sea mayor. Formalmente, la combinación producto se expresa de la siguiente forma:

$$Y = \arg máx_{j=1}^{c} \prod_{i=1}^{L} p_{i,j} \tag{7}$$

- **Combinación por mayoría de votos, o *majority voting***. Es una estrategia de combinación en la que solo se tiene en cuenta la clase más probable proporcionada por los clasificadores base y se elige la etiqueta más frecuente de entre dichos resultados. Generalmente, pueden darse tres escenarios diferentes: cuando todas las etiquetas son iguales, se da una votación por unanimidad; si más de la mitad de las etiquetas son de la misma clase, entonces se trata de una mayoría simple; por último, si una clase recibe la mayoría de los votos, pero no más de la mitad, se da la pluralidad. En el caso del clasificador propuesto en este trabajo, en el que solo se van a diferenciar dos clases, la regla de mayoría simple y pluralidad coinciden. Asumiendo que la salida de los clasificadores es un vector binario c-dimensional tal que $\left[d_{i,1}, \dots, d_{i,c}\right]^T \in \{0,1\}^c, i = 1, \dots, L$, el *majority voting* devuelve la clase w_k si se cumple que:

$$\sum_{i=1}^{L} d_{i,k} = max_{j=1}^{c} \sum_{i=1}^{L} d_{i,j} \tag{8}$$

En los casos en los que el combinador no puede determinar la clase ganadora con la confianza suficiente o cuando se produce un empate, se utilizará una nueva clase w_{c+1} para indicar que el resultado es ambiguo. Teniendo esto en cuenta, la función de decisión queda de la siguiente forma:

$$\begin{cases} w_k, & \text{si} \sum_{i=1}^{L} d_{i,k} \geq \alpha L \\ w_{c+1,} & \text{en cualquier otro caso} \end{cases} \tag{9}$$

donde α indica el grado de confianza necesario para tomar una decisión y es un valor comprendido entre 0 y 1. De esta forma, cuando α=1, la ecuación (9) se convierte en voto por unanimidad, es decir, que solo se asigna una clase únicamente si todos los clasificadores están de acuerdo.

3.6. Probabilidad acumulada

Una de las metas principales de este trabajo es proporcionar las bases de un sistema de apoyo a los médicos. Una característica deseable en este aspecto es poder evaluar periódicamente el riesgo de que un paciente desarrolle o no la enfermedad, desde un punto de vista estadístico. La probabilidad de detección y la probabilidad de detección acumulada [66] son dos métricas que resultan muy útiles para esta tarea. La probabilidad de detección, p_d, en una única oportunidad es la probabilidad de que el sistema detecte el evento de interés (p.ej., sepsis) en un instante de muestreo determinado, mientras que la probabilidad de detección acumulada, P_d, es la probabilidad de realizar una detección después de una serie de oportunidades consecutivas.

La probabilidad de detección acumulada se calcula en una ventana deslizante de tamaño K. Para cada instante de evaluación, n es el número mínimo de detecciones requerido para considerar que se hay riesgo de sepsis. La probabilidad de detección acumulada después de K oportunidades, $P_{d,K}$, se calcula de la siguiente forma:

$$P_{d,K} = 1 - \prod_{k=1}^{K'} (1 - p_{d,k}) \tag{10}$$

Donde K' $(n \leq K' \leq K)$ es el número de detecciones positivas obtenido en la ventana. Para el caso particular presentado en este trabajo, la probabilidad de detección singular, $p_{d,k}$, se corresponde con las probabilidades condicional $P\{D|T\}$ descrita en el apartado 3.3.2. Cuando la probabilidad $P_{d,K}$ supera un umbral determinado, se dispara una alerta de sepsis. Este umbral se estima durante el entrenamiento a partir de las curvas ROC (*Receiver Operating Characteristic*) generadas por el clasificador. Este concepto ha sido desarrollado con mayor profundidad en el apartado 3.8.1.

De la misma forma, si no se llega al número de detecciones mínimo establecido, se puede obtener la probabilidad de no detección acumulada, es decir, la probabilidad de que un paciente no vaya a desarrollar sepsis en la ventana de análisis. Esta probabilidad denomina $P_{d,\bar{K}}$, y se calcula de forma análoga a la ecuación (10):

$$P_{d,\bar{K}} = 1 - \prod_{k=1}^{\bar{K}} (1 - p_{d,k}) \quad (11)$$

Donde $\bar{K} = K - K'$ y $p_{d,k}$ es la probabilidad condicional $P\{\bar{D}|\bar{T}\}$. Si para un paciente nunca se llega al número de detecciones mínimas requeridas, la probabilidad que se acumula es $P_{d,\bar{K}}$ y no $P_{d,K}$, por lo que nunca se desatará la alerta. En cualquier caso, hay que destacar nuevamente que, en línea con el pensamiento científico generalizado, el que se presenta en este TFM pretende ser una ayuda para los médicos, y en ningún caso toma precedencia sobre el diagnóstico profesional.

3.7. Otros clasificadores

Para afirmar con la suficiente seguridad que el método propuesto en este trabajo es científicamente competitivo, es necesario completar el estudio del rendimiento de los clasificadores convencionales y, ulteriormente, comparar sus resultados con el clasificador Bayesiano. La regresión lineal, la regresión logística y la SVM son tres de las técnicas de Machine Learning no profundo más usadas para la detección de la sepsis, de acuerdo con una revisión sistemática [67]. A continuación, se describen brevemente los aspectos principales de estos métodos.

3.7.1. Regresión lineal

El objetivo de la regresión lineal es predecir el comportamiento de una variable dependiente Y y un conjunto de variables independientes $X_1, X_2 \ldots X_n$. En este TFM, $n = 1$, pues hay un clasificador base por cada variable clínica. Se puede decir que existe regresión de los valores de una variable con los de otra cuando hay alguna línea, denominada línea de regresión, que se ajusta en mayor o menor medida a los valores observados [68].

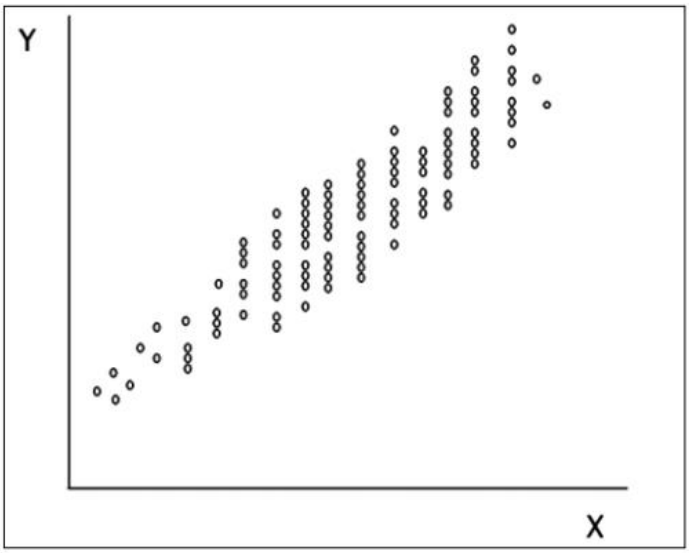

Figura 5. Nube de puntos. Fuente: [68].

Al representar los valores de ambas variables en un gráfico de coordenadas se puede obtener una "nube de puntos" (x_n, y_n) similar a la de la Figura 5. Las técnicas de regresión consisten en buscar una función que sea una buena aproximación de dicha nube de puntos. Mientras más dispersos estén los puntos, menos probable es que estén relacionados. Por otro lado, si los puntos están muy concentrados se puede asegurar que existe una relación determinista entre ambas variables [68].

Existen diferentes métodos para calcular la recta que mejor se ajuste a los puntos observados, pero el más usado consiste en medir la distancia vertical desde cada punto hasta la recta propuesta. Para ello se hace uso del método de los mínimos cuadrados [68] [69]. Este método consiste en minimizar la suma de los cuadrados de los errores, es decir, la diferencia entre los valores reales observados (y_n) y los valores estimados $(\hat{y}_n)$. En la Figura 6 se muestra un ejemplo de la recta estimada y el error ente valores reales y predichos para una regresión lineal simple.

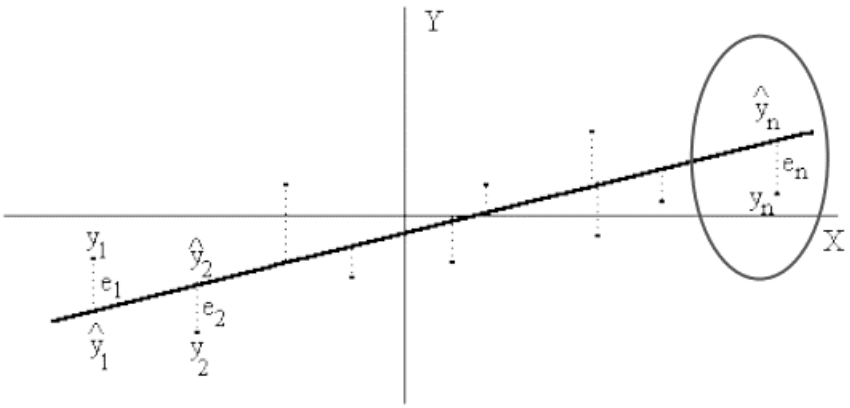

Figura 6. Error entre valores reales y valores estimados. Fuente: [69].

La ecuación obtenida será de la siguiente forma:

$$\hat{Y} = \alpha + \beta X = a + bX \qquad (1)$$

Donde α es el punto de intersección y β la pendiente de la recta predicha. Los valores a y b se denominan coeficientes de la recta de regresión de los cuadrados mínimos, y son las mejores estimaciones de α y β. El valor de a generalmente carece de importancia, mientras que el signo de b aportará información acerca de si las variables son directamente proporcionales (signo positivo), o por el contrario se comportan de forma inversa (signo negativo). Por último, se usa la varianza para obtener el grado de variabilidad de los datos alrededor de la línea de regresión. Cuando la varianza es cero, todos los puntos Y coinciden con la recta, y la relación entre las variables es determinista. Por otro lado, cuanto mayor es la varianza menos se puede usar X para predecir Y debido a que aumenta la incertidumbre. El cálculo de la varianza se hace con la distancia de cada punto hasta la media y el tamaño de la muestra:

$$s^2 = \frac{1}{n-2}\sum_i d_i^2 = \frac{1}{n-2}\sum_i (y_i - \hat{y}_n) \qquad (2)$$

La regresión lineal no puede aplicarse sobre cualquier tipo de variable. Generalmente, cuando una variable es continua, el modelo de regresión lineal es el más utilizado. Sin embargo, cuando la variable de interés es dicotómica, es más adecuado emplear un modelo de regresión logística [70].

3.7.2. Regresión logística

Cuando se aplica un modelo de regresión logística, en lugar de estimar los valores reales de la variable de interés se obtiene una función basada en la probabilidad de que

dicha variable adopte el valor previamente definido. Al igual que en el apartado anterior, $n = 1$ pues hay un clasificador base por cada variable clínica.

La función obtenida tiene la siguiente forma:

$$Y = \ln\left(\frac{p}{1-p}\right) \tag{14}$$

Esto implica que la variable estimada puede, en principio, tomar cualquier valor, al contrario de la regresión lineal donde estaba restringida a un rango de valores determinado. Posteriormente, se pueden usar los métodos de regresión aplicables al modelo lineal sobre esta probabilidad. En general, ambos modelos se diferencian únicamente en la interpretación de los resultados obtenidos. En el caso de la regresión logística, se emplea el concepto "*odds"* [70], esto es, el cociente entre la probabilidad de presentar una característica o no presentarla. Comparando los *odds* de una característica concreta entre pacientes con una misma enfermedad es posible, por ejemplo, concluir si la enfermedad es más común en pacientes en los que se da esta característica o no. El cociente entre los *odds* se denomina "*odds ratio"*.

$$\exp(\beta) = odds\ ratio \tag{15}$$

Los coeficientes obtenidos en el proceso de regresión logística (β) representan el riesgo de presentar cierta característica respecto a no presentarla, de forma que se puede obtener el *odds ratio* con la igualdad (15). Si la variable independiente estudiada es numérica, β se interpreta como la variación en el riesgo cuando se incrementa el valor de la variable en una unidad manteniendo el resto de las variables constantes [70].

Como se ha comentado anteriormente, la regresión logística se encuentra entre los métodos de Machine Learning más comunes aplicados al problema de detección y predicción de la sepsis. En este aspecto, distintos estudios [71] [72] han concluido que la capacidad predictiva de un modelo basado en regresión logística es comparable a la de una red neuronal.

3.7.3. Support Vector Machine

Support Vector Machine es un algoritmo de aprendizaje supervisado usado en tareas de clasificación y regresión. El objetivo principal de SVM es encontrar un hiperplano de clasificación que divida dos clases de datos de forma óptima. En un espacio de n dimensiones, un hiperplano es un subconjunto plano de $n-1$ dimensiones que divide el espacio en dos partes completamente separadas [73]. Además de dividir las muestras

correctamente, el hiperplano óptimo es aquel que maximiza la distancia o margen entre clases [74].

El funcionamiento de SVM se puede entender a través de un ejemplo [73]. En la Figura 7 se muestra un conjunto de datos de dos clases diferentes, denominadas clase A y clase B. La complejidad radica en el hecho de que existen infinitas líneas que separan ambas clases, por lo que el algoritmo debe determinar cuál es la mejor opción para maximizar la eficiencia de la clasificación y minimizar los errores. Para ello, primero identifica los puntos de cada clase más cercanos a las líneas, denominados vectores de soporte. Posteriormente, se calcula la distancia entre la línea y los vectores de soporte, es decir, el margen. El hiperplano máximo será aquel para el que este margen es óptimo, como se muestra en la Figura 8.

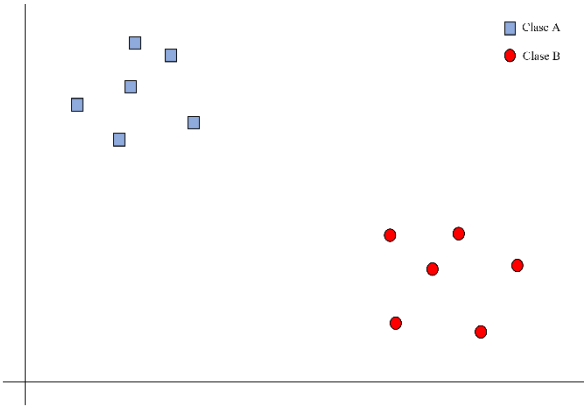

Figura 7. Dos clases distribuidas en un espacio. Adaptado de [73].

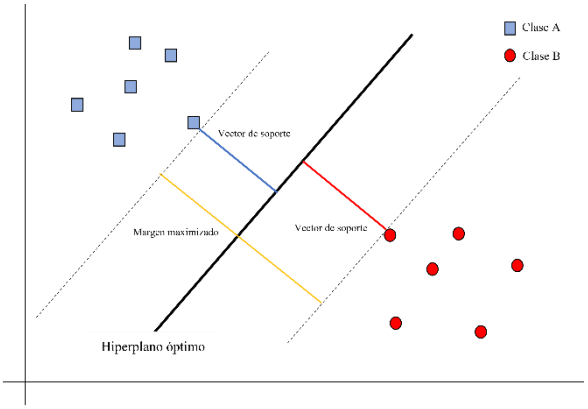

Figura 8. Hiperplano óptimo calculado por SVM. Adaptado de [73].

El algoritmo calcula el margen óptimo en problemas que permiten la separación lineal de los datos, que es un caso ideal. En la práctica, cuando los datos no se puedan separar

linealmente, el algoritmo intentará maximizar el margen a costa de un pequeño número de errores de clasificación [75].

SVM ha sido ampliamente probado en el campo de detección de la sepsis, y ha demostrado una mayor capacidad predictiva frente a otros métodos populares como la regresión lineal o el árbol de modelos logísticos [76]. Al igual que en los apartados anteriores, $n - 1$ pues hay un clasificador base por cada variable clínica.

3.8. Métricas de calidad

Cuando se diseñan metodologías y algoritmos de clasificación aplicados a la monitorización de pacientes es necesario evaluar su calidad. Dicha evaluación permite estudiar si se alcanzan los objetivos que se habían marcado, así como comparar con otros algoritmos y verificar si es científicamente competitivo. Es muy común emplear métricas de calidad para dicha evaluación de un sistema basado en Machine Learning. Para el trabajo presentado en este TFM, los principales medios que se han usado para la evaluación de los algoritmos propuestos han sido las curvas ROC, la matriz de confusión y métricas derivadas (sensibilidad, especificidad, etc.), y la función de utilidad [14] del reto PhysioNet. Esto se verá con más detalle a continuación.

3.8.1. Curvas ROC

El término "Característica Operativa del Receptor" surgió en el campo de detección de señales como una forma de diferenciar las señales verdaderas de ruido. Debido al paralelismo entre el solapamiento de señales con ruido y el solapamiento entre los resultados de pacientes con una enfermedad y aquellos que no la tienen, estas técnicas se comenzaron a aplicar a los sistemas de diagnóstico médico [77] [78]. La curva ROC es una representación gráfica de la sensibilidad frente a la especificidad de un sistema de clasificación binario. Cada punto de la curva ROC representa una pareja de sensibilidad y especificidad, asociado a un umbral de decisión determinado. En el eje de Y se sitúa la sensibilidad, o fracción de verdaderos positivos, mientras que en el eje X se sitúa la fracción de falsos positivos, es decir, 1 – especificidad. En la Figura 9 se muestra un ejemplo de curva ROC simple.

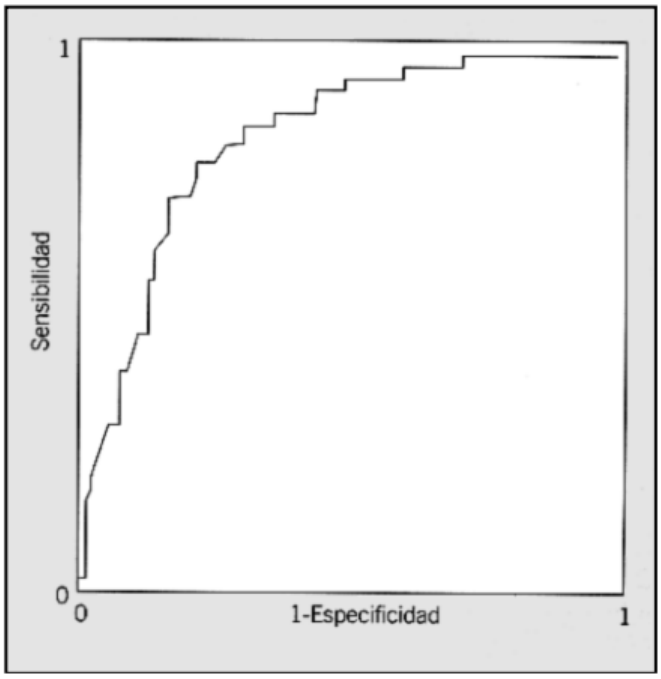

Figura 9. Ejemplo de una curva ROC. Fuente: [78].

El área bajo la curva ROC (*Area Under the ROC curve*, AUROC) representa el valor de la capacidad de discriminación de un test diagnóstico. En la Tabla 11 se muestra una relación entre el valor del área y la exactitud esperada del test [79]. Como se puede apreciar, la exactitud del test aumenta junto con el área. Se considera que un test con un AUROC por debajo de 0,7 tiene una capacidad de predicción mala, si bien esta es una valoración subjetiva que dependerá en gran medida de la aplicación de que se trate.

Tabla 11. Interpretación de la exactitud del test basada en el AUROC. Fuente: [79]

AUROC	Interpretación de la exactitud del test
1	Perfecta
0,9-1	Excelente
0,8-0,9	Buena
0,7-0,8	Aceptable
0,5-0,7	Mala

Las curvas ROC se emplean para cuatro propósitos principales [79]: determinar el umbral de decisión para el cual el número de predicciones incorrectas es mínimo; evaluar la habilidad discriminante del test; comparar la habilidad discriminante de dos o más test

de diagnóstico de la misma enfermedad; y comparar dos o más observadores que están realizando la misma prueba (variabilidad entre observadores).

A la hora de elegir el umbral óptimo, se ha optado por el criterio ampliamente utilizado en la comunidad científica de elegir aquel que dé el mejor compromiso entre máxima sensibilidad y máxima especificidad. Gráficamente es el umbral asociado a la curva ROC más cercana al extremo superior izquierdo donde la sensibilidad es 1 y (1-especificidad) es 0. No obstante, podrían aplicarse otros criterios. Por ejemplo, si las omisiones de casos de sepsis son particularmente críticas (interesa una alta probabilidad de detección), podría primarse mantener la probabilidad de omisión por debajo de un umbral, aún a costa de incrementar las falsas alarmas. Esto sería objeto de discusión interdisciplinar en los ámbitos científico-técnico y científico-médico. En cualquier caso, el criterio elegido no condiciona la metodología usada, aunque sí los resultados.

3.8.2. Matriz de confusión

Cuando se tratan problemas de clasificación, los errores cometidos por el algoritmo representan una importante fuente de información, ya que permiten identificar problemas en el modelo y mejorar su rendimiento realizando los ajustes necesarios. Generalmente, cuando se emplean técnicas de aprendizaje supervisado, en las que los datos están etiquetados previamente y se conoce la distribución de las diferentes clases, el principal método de evaluación del modelo es la matriz de confusión.

Tabla 12. Matriz de confusión para una clasificación binaria.

		Valor predicho	
		Positivo	Negativo
Valor real	Positivo	TP (*True Positive*)	FN (*False Negative*)
	Negativo	FP (*False Positive*)	TN (*True Negative*)

Se pueden diferenciar tres tipos de matrices: binaria, cuando cada etiqueta puede tomar solo dos valores; multiclase, cuando cada etiqueta puede tomar más de dos valores; y múltiples etiquetas, cuando cada observación está asociada a múltiples clases. La matriz de confusión para una clasificación binaria, que es el caso de este trabajo, se muestra en

la Tabla 12. Como se puede observar, la distribución de la matriz es similar a la presentada en la Tabla 10. Sin embargo, en este caso no se trata de probabilidades condicionales, sino que se representa el número de predicciones de cada tipo que ha hecho el clasificador.

Cada columna de la matriz representa las predicciones del clasificador, mientras que cada fila representa las clases reales. Un elemento de la matriz de confusión en la fila *i* y columna *j* indica el número de casos en el que la clase real es *j* y se ha clasificado como clase *i* [80]. Así, TP (*True Positive*) representa el número de veces que la clase positiva se ha clasificado correctamente, mientras que FN (*False Negative*) es el número de veces que el clasificador ha predicho erróneamente que un caso positivo es negativo. Así, la matriz de confusión no solo proporciona información sobre los errores que comete un clasificador, sino que además permite identificar la naturaleza del error cometido.

A continuación, se describen las principales métricas de calidad que se pueden derivar a partir de los datos de la matriz de confusión.

3.8.3. Exactitud

La exactitud hace referencia al porcentaje de aciertos del test. Se expresa como la relación entre el número de predicciones correctas y el número total de casos. Para el caso particular de una clasificación binaria, como es el caso de este trabajo, la exactitud se calcula con la siguiente igualdad:

$$Exactitud = 100 \cdot \frac{TP + TN}{TP + FN + TN + FP} \qquad (16)$$

Un alto porcentaje de exactitud indica que la predicción realizada por el modelo se ajusta a la realidad de los datos. No obstante, la exactitud puede llevar a error cuando se está trabajando con una base de datos desbalanceada. Supóngase, por ejemplo, que el 80% de los pacientes no presentan enfermedad y que el modelo solo predice correctamente esta porción de los datos. La exactitud calculada será del 80%, pero el modelo no es preciso, ya que no ha sido capaz de tratar la clase minoritaria, que es, generalmente, la más delicada. Debido a esto, no es recomendable usar únicamente la exactitud como única medida de la bondad del test.

3.8.4. Sensibilidad

La sensibilidad es la tasa de verdaderos positivos del test, es decir, el porcentaje de aciertos que ha tenido el modelo a la hora de identificar un paciente con sepsis.

Matemáticamente, la sensibilidad se expresa como la relación entre los verdaderos positivos predichos y los positivos totales:

$$Sensibilidad = 100 \cdot \frac{TP}{TP + FN} \quad (17)$$

Cuanto mayor es la sensibilidad del test, más casos positivos son detectados. Por tanto, lo deseable es que el valor de esta métrica sea lo más elevado posible.

3.8.5. Especificidad

La especificidad es la tasa de verdaderos negativos del test. Dicho de otra forma, se trata del porcentaje de pacientes sin sepsis que han sido clasificados correctamente como negativos por el algoritmo. De forma análoga a la sensibilidad, la especificidad es la relación entre los verdaderos negativos predichos y los casos negativos totales:

$$Especificidad = 100 \cdot \frac{TNP}{TN + FP} \quad (18)$$

Una baja especificidad implica un elevado número de falsas alarmas. Un test ideal tendría alta especificidad y sensibilidad, sin embargo, esto no se suele dar en la práctica. Por lo general, se hace necesario un compromiso entre ambas métricas, de forma que cuando se optimiza una de ellas la otra empeora notablemente. Por tanto, es necesario tener en cuentas las necesidades del problema que se está abordando para tomar una decisión en cuanto a qué aspecto del test es más prioritario: alta sensibilidad o alta especificidad.

Existen tres escenarios clínicos en los que se requiere una alta sensibilidad [61]: cuando existe una gran penalización por no detectar un paciente con la enfermedad, cuando la probabilidad de tener la enfermedad es baja y el único propósito del test es descubrir individuos asintomáticos, y en las primeras etapas del desarrollo de una enfermedad. Asimismo, se pueden considerar dos escenarios principales en los que se prefiere un test con alta especificidad [61]: cuando un falso positivo puede causar daños emocionales o físicos al paciente, o cuando se pretende descartar un diagnóstico sugerido por otros test.

3.8.6. Función de utilidad

Como se había comentado anteriormente, además de la evaluación de los algoritmos mediante las métricas estándar ya mencionadas, en el reto propuesto [14] se introdujo una función de utilidad especialmente creada dicha ocasión. La función de utilidad premia a los clasificadores que predicen la sepsis precozmente y penaliza las predicciones

tempranas erróneas o las tardías. Asimismo, penaliza la detección de sepsis en pacientes no sépticos.

En el conjunto de datos se definen las siguientes marcas temporales para cada paciente:

- $\boldsymbol{t_{suspicion}}$: sospecha clínica de infección, identificada como el instante más temprano de administración de antibióticos intravenosos y hemocultivos dentro de una duración especificada. Si se administraron primero los antibióticos, entonces los cultivos deben haberse obtenido en un plazo de 24 horas. Si se obtuvieron primero los cultivos, entonces el antibiótico debe haber sido ordenado en las 72 horas siguientes. Los antibióticos deben haberse administrado durante 72 horas seguidas para ser considerados.
- $\boldsymbol{t_{SOFA}}$**:** aparición de daño en algún órgano importante, identificado por un deterioro de dos puntos de la puntuación SOFA en un período de 24 horas.
- $\boldsymbol{t_{sepsis}}$**:** inicio de la sepsis. Es el instante más temprano entre $t_{suspicion}$ y t_{SOFA}, siempre que t_{SOFA} no se dé más de 24 horas antes o 12 horas después de $t_{suspicion}$. Si no se cumple este requisito, el paciente no se marcará como séptico. Dicho de otra forma, si $t_{suspicion} - 24 \leq t_{SOFA} \leq t_{suspicion} + 12$, entonces $t_{sepsis} =$ mín $(t_{suspicion}, t_{SOFA})$.

Un algoritmo realiza una predicción binaria para cada intervalo de tiempo t en cada uno de los archivos de paciente s, por lo que cada predicción obtiene una puntuación de utilidad $U(s,t)$. De acuerdo con la ecuación (19), la puntuación final del algoritmo para un conjunto de datos es la suma de las puntuaciones de todos los intervalos de tiempo para todos los pacientes.

$$U_{total} = \sum_{s \in S} \sum_{t \in T(s)} U(s,t) \qquad (19)$$

En función de si el paciente analizado tiene sepsis eventualmente o no, se recompensa o penaliza al algoritmo de forma diferenciada. El momento en que se produce la detección de la sepsis se denomina t_{sepsis}, y se representa con un 1 en la etiqueta 'SepsisLabel' del paciente. Para los pacientes que desarrollan sepsis, el algoritmo será premiado si predice la sepsis entre 12 y 3 horas antes de t_{sepsis}, siendo 1,0 la recompensa máxima que se puede obtener. Por otro lado, se penaliza a los algoritmos que no llegan a predecir la sepsis o que la predicen más de 12 horas antes del momento de detección clínica. En este

caso, la penalización máxima percibida por el algoritmo será de 0,05 y 2,0, respectivamente. Por el contrario, para un paciente que no tiene sepsis, no se recompensa ni se castiga a los algoritmos que no predicen la sepsis, pero sí que se penalizan las falsas alarmas. En este caso, se reducirá un máximo de 0,05 a la puntuación por cada error, la misma penalización que se aplica a una detección muy temprana de la sepsis. En la Figura 10 se muestra dos representaciones de la asignación de puntuaciones dependiendo del momento de detección de la sepsis.

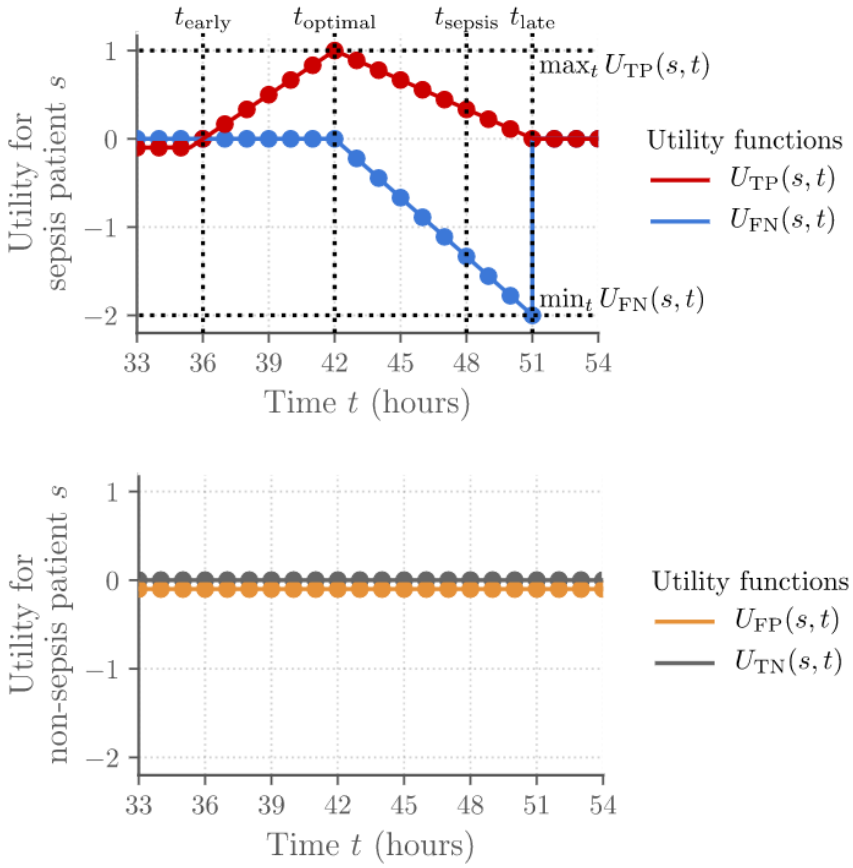

Figura 10. Función de utilidad (A: pacientes con sepsis; B: pacientes sin sepsis). Fuente: [81].

Como se aprecia en la figura anterior, con el fin de mejorar la interpretabilidad del resultado se normaliza la puntuación total U_{total} de forma que un clasificador óptimo recibe la máxima puntuación de 1, mientras que un clasificador pasivo, es decir, que no realiza predicciones positivas, recibirá una puntuación normalizada de 0:

$$U_{normalizada} = \frac{U_{total} - U_{no\ predicciones}}{U_{optima} - U_{no\ predicciones}} \tag{20}$$

Capítulo 4. Conjunto de datos

El conjunto de datos que se ha usado ha sido el proporcionado para el reto "Early Prediction of Sepsis from Clinical Data - The PhysioNet Computing in Cardiology Challenge 2019" [14]. Dicho conjunto está formado por dos subconjuntos diferentes, denominados "training_setA" y "training_setB". Ambos subconjuntos contienen datos de pacientes internados en la UCI de dos sistemas hospitalarios diferentes. Los organizadores del reto contaron con un tercer subconjunto de datos adicional, el "training_setC", que se usó para la evaluación de los participantes del reto. Dado que no se ha podido tener acceso al "training_setC", se trabajará con los conjuntos de datos A y B.

Cada registro de paciente está contenido en un único archivo de texto divididos en columnas, donde cada fila representa los datos recogidos con periodicidad de una hora entre muestras consecutivas. Para cada paciente se incluyen datos demográficos, signos vitales y valores de laboratorio.

4.1. Estructura de los datos

Cada fichero de paciente contiene el mismo encabezado, que contiene el nombre de las variables disponibles. Los datos están dispuestos en forma de tabla, en la que las columnas representan cada variable y las filas los datos recogidos una hora determinada. Por tanto, todos los pacientes tendrán el mismo número de columnas, pero distinto número de filas dependiendo del tiempo que hayan pasado ingresados. En total, se dispone de 41 variables que se dividen en signos vitales (Tabla 13), valores de laboratorio (Tabla 14), datos demográficos (Tabla 15) y resultado (

Tabla 16), es decir, si el paciente es séptico o no.

Tabla 13. Columnas 1-8 en cada fichero de paciente. Fuente: [81]

Signos vitales		
Nombre	***Descripción***	***Unidades***
HR	Frecuencia cardiaca	latidos/min
O2Sat	Saturación de oxígeno	%
Temp	Temperatura	°C
SBP	Presión arterial sistólica	mmHg
MAP	Presión arterial media	mmHg

DBP	Presión arterial diastólica	mmHg
Resp	Frecuencia respiratoria	respiraciones/min
EtCO2	Dióxido de carbono espiratorio final	mmHg

Tabla 14. Columnas 9-34 de cada fichero de paciente. Fuente: [81]

Valores de laboratorio		
Nombre	***Descripción***	***Unidades***
BaseExcess	Exceso de bicarbonato	mmol/L
HCO3	Bicarbonato	(0) mmol/L
FiO2	Fracción de oxígeno inspirado	%
pH	pH	-
PaCO2	Presión parcial de CO2 en la sangre arterial	mmHg
SaO2	Saturación de oxígeno en la sangre arterial	%
AST	Aspartato trasaminasa	IU/L
BUN	Nitrógeno ureico en sangre	mg/dL
Alkalinephos	Fosfata alcalina	IU/L
Calcium	Calcio	mg/dL
Chloride	Cloro	mmol/L
Creatinine	Creatinina	mg/dL
Bilirubin_direct	Bilirrubina directa	mg/dL
Glucose	Glucosa	mg/dL
Lactate	Lactato	mg/dL
Magnesium	Magnesio	mmol/L
Phosphate	Fosfato	mg/dL
Bilirubin_total	Bilirrubina total	mg/dL
Troponin I	Troponina I	ng/mL
Hct	Hematocrito	%
Hgp	Hemoglobina	g/dL
PTT	Tiempo de tromboplastina parcial	s
WBC	Recuento de leucocitos	count/mL
Fibrinogen	Fibrinógeno	mg/dL
Platelets	Recuento de plaquetas	count/mL
BaseExcess	Exceso de bicarbonato	mmol/L

Tabla 15. Columnas 35-40 de cada fichero de paciente. Fuente: [81]

Datos demográficos		
Nombre	***Descripción***	***Unidades***
Age	Edad	Años (100 para pacientes de 90 años en adelante)
Gender	Género	(1) - Femenino (2) - Masculino
Unit1	Identificador administrativo de la unidad de la UCI (MICU)	-
Unit2	Identificador administrativo de la unidad de la UCI (SICU)	-
HospAdmTime	Horas entre la admisión en el hospital y la admisión en la UCI	Horas
ICULOS	Tiempo de estancia en la UCI	Horas

Tabla 16. Columna 41 de cada fichero de paciente. Fuente: [81]

Resultado		
Nombre	***Descripción***	***Unidades***
SepsisLabel	Etiqueta de sepsis	Para pacientes con sepsis, 1 si $t \geq t_{sepsis} - 6$ y 0 si $t < t_{sepsis} - 6$. Para pacientes sin sepsis, 0.

4.2. Características generales de los datos

En total, el conjunto de datos consta de 40.336 pacientes, de los cuales 20.336 pertenecen al subconjunto A, mientras que los 20.000 pacientes restantes pertenecen al subconjunto B. En ambos conjuntos hay datos de pacientes sépticos y no sépticos. Hay que destacar que la proporción de pacientes de ambas clases no es la misma y hay un gran desequilibrio en el número de pacientes, es decir, las clases están desbalanceadas. En la Figura 11 se muestra la relación de pacientes de cada clase en ambos conjuntos de datos.

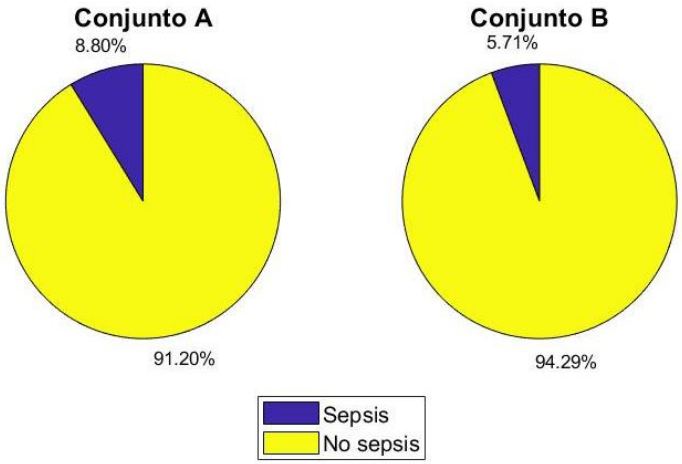

Figura 11. Porcentaje de pacientes con y sin sepsis en el conjunto de datos (Izq: Conjunto A. Dcha: Conjunto B).

Como se observa en la figura superior, solo el 8,80% de los pacientes contenidos en el conjunto A son pacientes que desarrollaron sepsis, frente al 91,20% de pacientes sanos. En el caso del subconjunto B esta diferencia es aún mayor, ya que únicamente un 5,71% de los datos pertenecen a pacientes sépticos. Si en lugar de contabilizar los pacientes se analizan las horas con y sin sepsis, el 2,17% de las horas en el conjunto A son horas con sepsis, mientras que en el conjunto B hay un 1,41% de horas con sepsis, como se observa en la Figura 12.

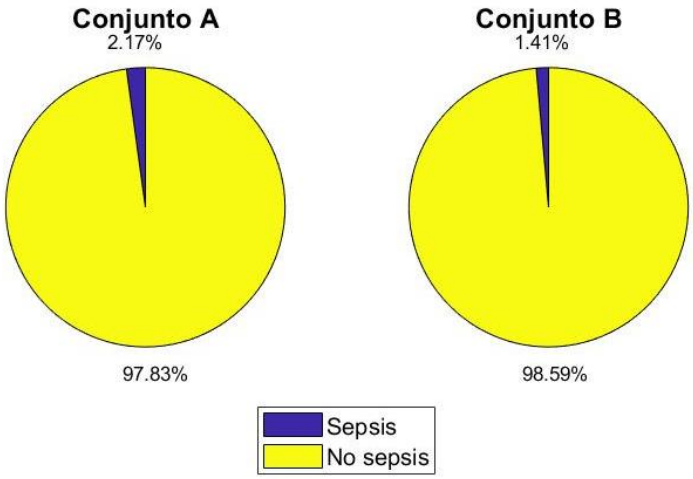

Figura 12. Porcentaje de horas con y sin sepsis en el conjunto de datos (Izq: Conjunto A. Dcha: Conjunto B).

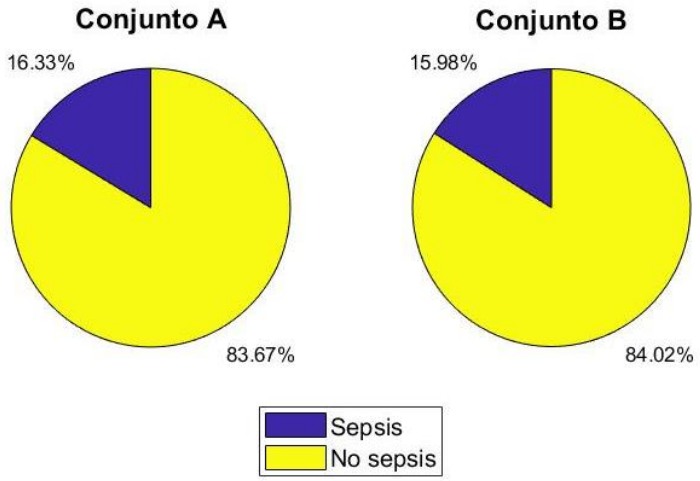

Figura 13. Porcentaje de horas con y sin sepsis en pacientes con sepsis (Izq: Conjunto A. Dcha: Conjunto B).

Por último, si se analizan únicamente los pacientes con sepsis, solo el 16,33% de las horas de los pacientes con sepsis contenidos en el conjunto A han sido diagnosticadas como positivas, mientras que el 83.67% restante son horas sin sepsis o normales. Para el conjunto de datos B la proporción es muy similar, con el 15,98% de horas sépticas y el resto normales. Estos datos se muestran en la Figura 13.

Como se comentó anteriormente, el número de horas de estancia en la UCI varía entre pacientes. Esto puede deberse a lo rápido que mejore o empeore un paciente y a las sospechas médicas de que vaya a desarrollar la enfermedad. Por ejemplo, en la Figura 14 se muestra el número de horas ingresados en la UCI de los pacientes con sepsis contenidos en el subconjunto A, mientras que en la Figura 15 se muestran los mismos datos para los pacientes no sépticos. Se puede apreciar a simple vista que en el caso de los pacientes que no llegaron a desarrollar sepsis el número de horas que estuvieron ingresados es mucho menor en general que para el grupo opuesto. Así, ninguno de los pacientes no sépticos llegó a superar las 100 horas ingresado, en contraposición a los pacientes que sí tuvieron la enfermedad, donde hay incluso casos de más de 300 horas en la UCI.

Por otro lado, en el subconjunto B esta diferencia de horas no es tan evidente. En las Figura 16 y 17 se muestran las horas de estancia en la UCI para pacientes sépticos y no sépticos, respectivamente. En este caso, hay un pequeño número de casos de pacientes sin sepsis que estuvieron entre 100 y 350 horas en la UCI. No obstante, la Figura 17 presenta una distribución bimodal, ya que la mayor parte de los pacientes están

concentrados alrededor de dos picos principales, en 15 y 30 horas aproximadamente, mientras que en el caso mostrado en la Figura 16 se puede observar una distribución de tipo exponencial con una cola muy alargada hacia las estancias prolongadas. Por tanto, si bien hay excepciones, en el subconjunto B también se cumple que los pacientes no sépticos tienen menos horas que los pacientes con sepsis, es decir, los pacientes sépticos tienen una mayor cantidad de medidas disponibles.

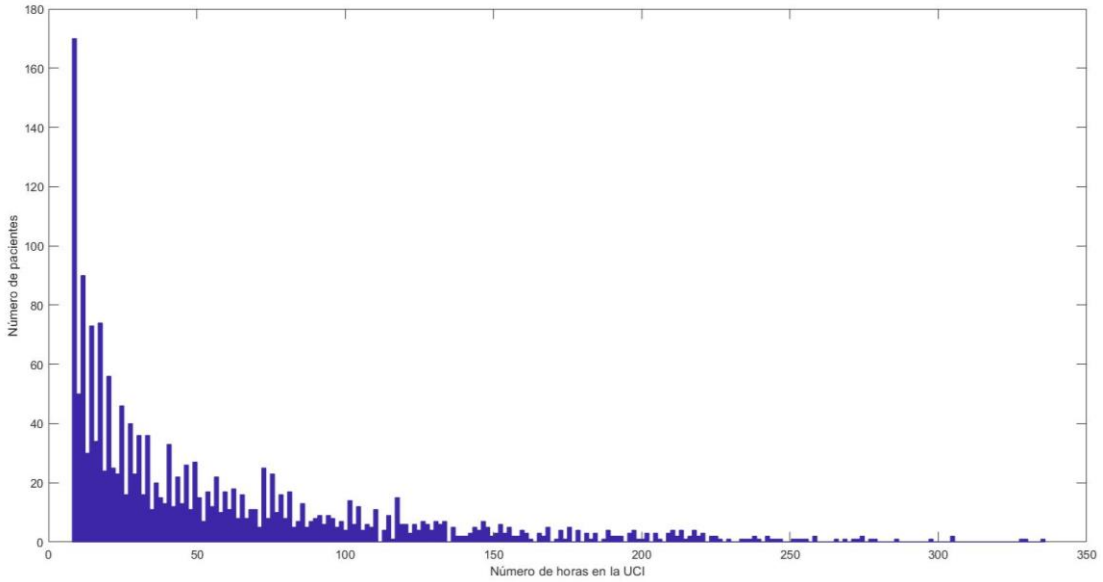

Figura 14. Número de horas en la UCI para pacientes sépticos en el conjunto de datos A.

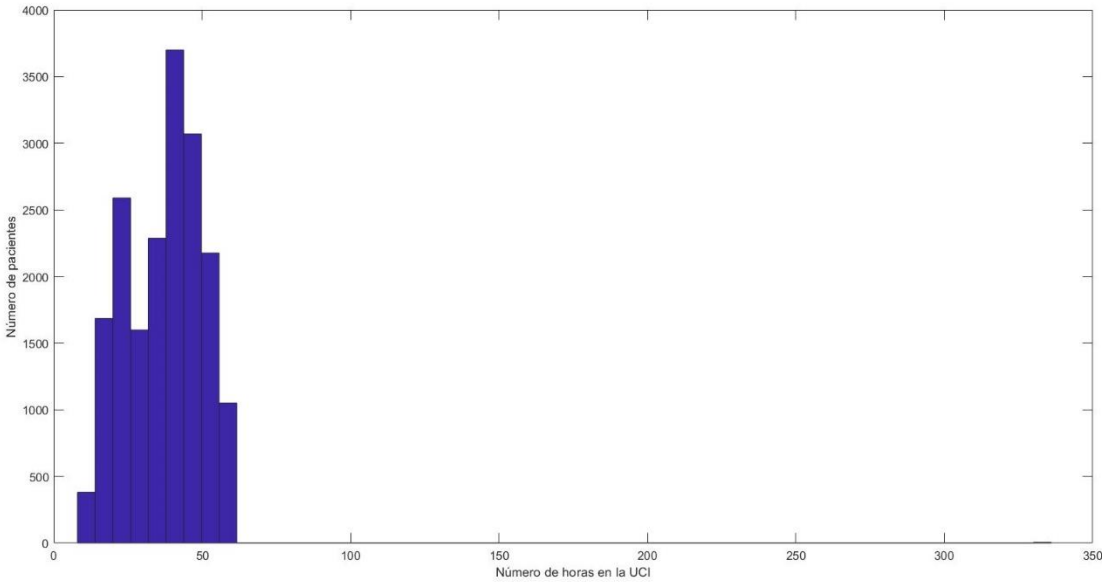

Figura 15. Número de horas en la UCI para pacientes no sépticos en el conjunto de datos A.

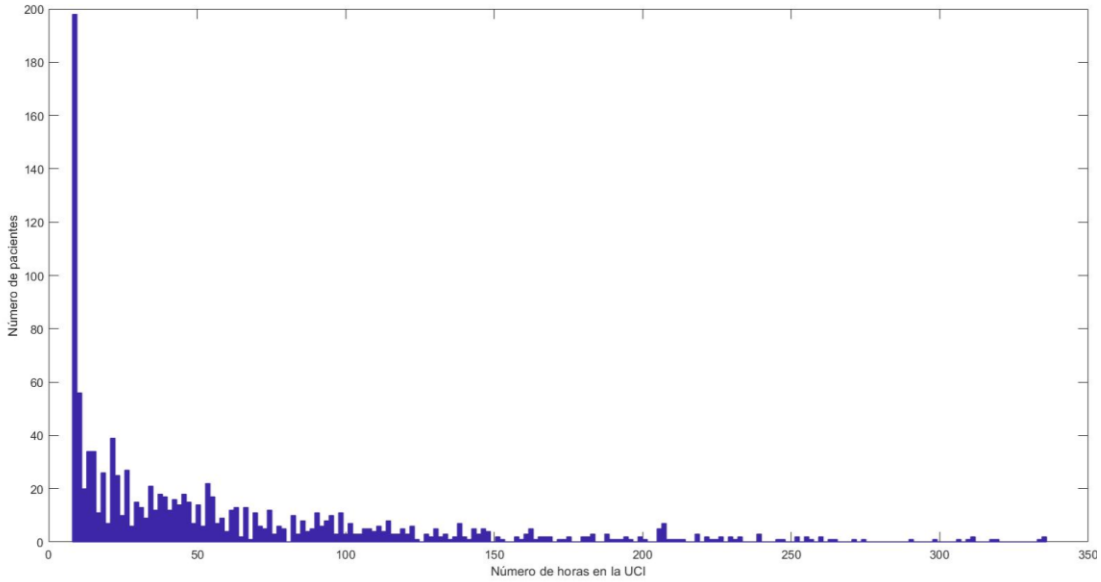

Figura 16. Número de horas en la UCI para pacientes sépticos en el conjunto de datos B.

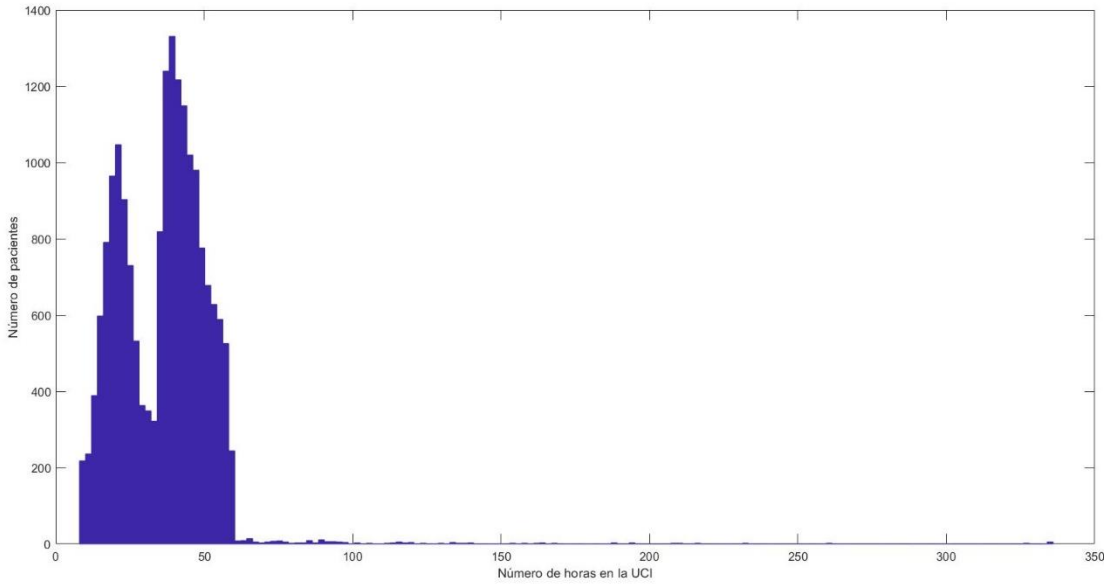

Figura 17. Número de horas en la UCI para pacientes no sépticos en el conjunto de datos B.

4.3. Datos perdidos

Si se analizan los registros de pacientes, se puede apreciar que hay instantes en los que las medidas de una o más variables clínicas tienen valor 'NaN' en lugar de un valor numérico. Esto indica que no se tiene la medida, es decir, es un dato perdido. En las Figura 18 y 19 se muestran los porcentajes de datos perdidos para las distintas variables en los conjuntos de datos A y B, respectivamente.

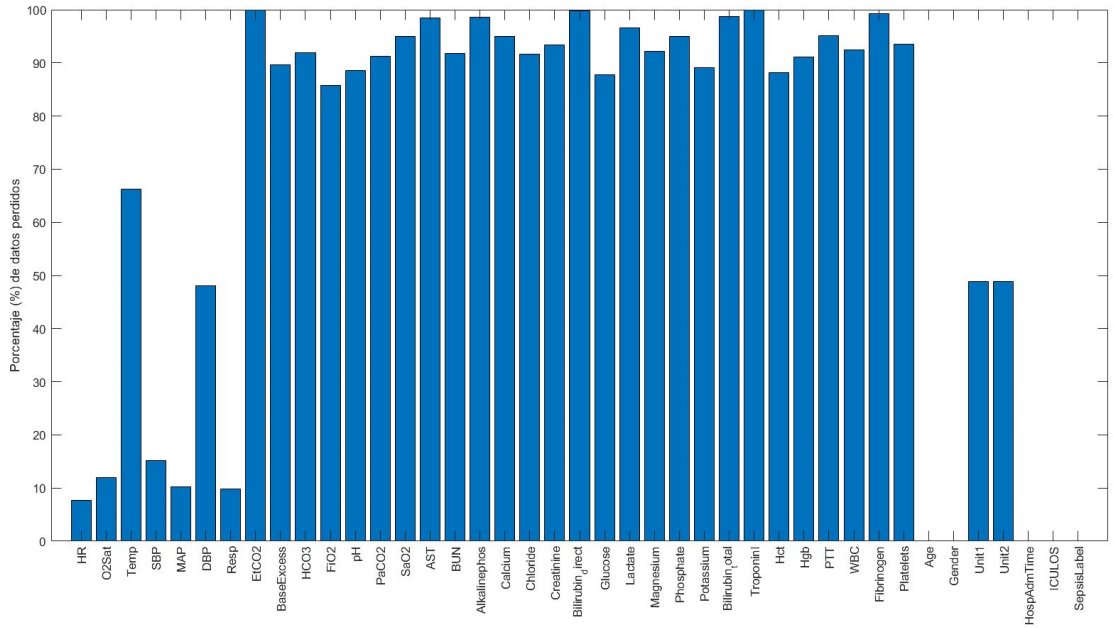

Figura 18. Porcentaje de valores perdidos en el conjunto de datos A.

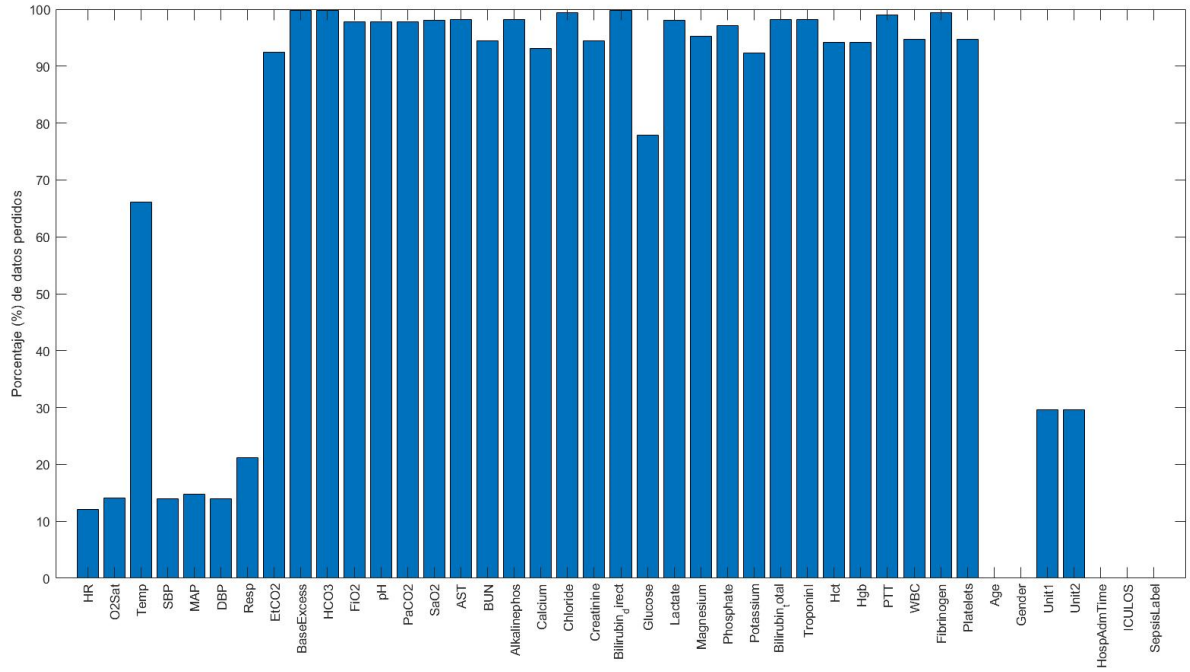

Figura 19. Porcentaje de valores perdidos en el conjunto de datos B.

Como se observa, la temperatura y ‘EtCO2’ presentan un alto porcentaje de valores perdidos junto con las variables de laboratorio (Tabla 14). Esto puede deberse a múltiples razones, como el orden irregular a la hora de realizar test de laboratorio y comprobar los signos vitales, o simplemente a que no todas las variables se recogen a intervalos regulares. Los variables demográficas no presentan datos perdidos, es decir, tienen un

100% de datos útiles. Esto es lo esperado, ya que esta información se toma una única vez en el momento del ingreso del paciente (como 'Age' o 'Gender'), o bien una vez que termina su estancia en la UCI (como 'ICULOS'). Sin embargo, es precisamente por esto que la utilidad de dichas variables es limitada, ya que no permiten analizar la evolución del paciente. En el caso de 'Unit1' y 'Unit2', debido a que son también valores constantes, si se conoce su valor presentarán un porcentaje de datos útiles del 100%. Por el contrario, si se desconocen estos identificadores su valor será 'NaN' en todas las entradas del registro, por lo que el porcentaje de datos perdidos será del 100%.

Conocer el porcentaje de datos perdidos que presenta cada variable es fundamental, ya que es uno de los principales criterios para seleccionar cuáles se van a usar para entrenar el modelo. Por ejemplo, las variables 'BaseExcess' y 'HCO3' tienen el 100% de datos perdidos en el conjunto de datos B. Si se planteara una estrategia de aprendizaje basada en validación cruzada con ambos subconjuntos, no se podrían emplear estas dos variables clínicas. Consecuentemente, solo se tendrán en cuenta aquellas variables que presenten un mínimo de datos útiles en los dos conjuntos.

4.4. Valores atípicos y aberrantes

Además de valores perdidos, se pueden encontrar tanto valores estadísticamente atípicos como *outliers* de una variable. Un *outlier* en un conjunto de datos es una observación cuyo valor está muy alejado de los valores que toman las observaciones, esto es, los típicos. Una forma sencilla de evaluar los *outliers* de una distribución es mediante el rango intercuartílico (*Interquartile Range,* IQR). IQR puede entenderse, de forma simplificada, como el rango en el que se encuentran la mitad de los valores de una distribución, y es la distancia entre dos cuartiles, Q_3 y Q_1:

$$\text{IQR} = Q_3 - Q_1 \tag{21}$$

Donde Q_1 es el estadístico por debajo del cual se encuentran el 25% de los valores y Q_3 es el estadístico por debajo del cual se encuentran el 75% de los valores [74]. En la Figura 20 se muestran los *outliers* de la variable 'HeartRate' en relación con IQR tanto para pacientes sépticos como no sépticos, en los dos conjuntos de datos.

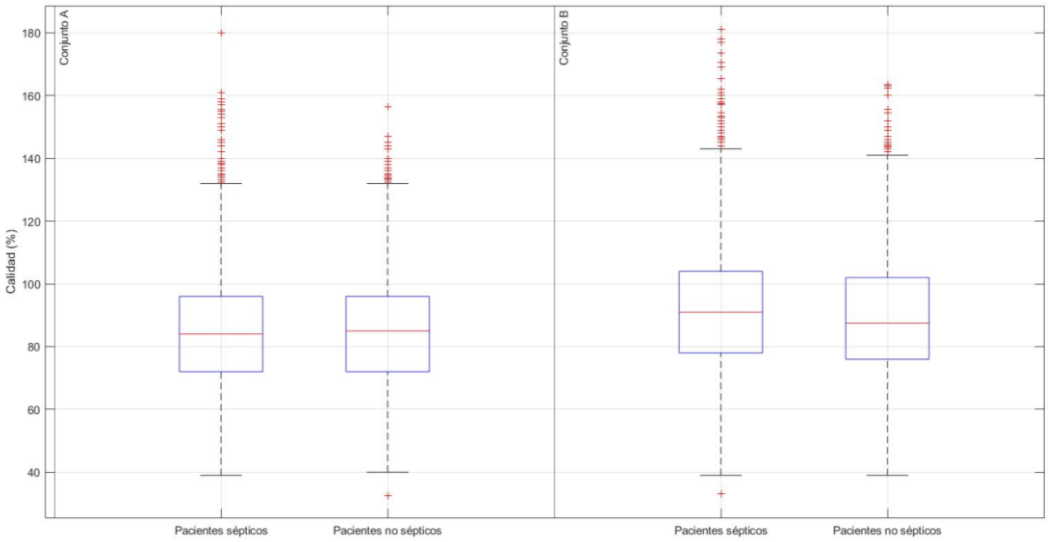

Figura 20. Valores outliers de la variable 'HeartRate'.

Por otro lado, en el contexto de una variable clínica, un valor aberrante puede identificarse como un valor imposible para dicha medida. Generalmente, estos valores son el producto de un error. Hay distintos tipos de error: durante la medida; a la hora de copiar o transcribir los datos manualmente; o porque una unidad que no es parte de la población objetivo se introduce de alguna forma en la muestra [82]. Existen distintos métodos para tratar los valores atípicos y *outliers*. En este TFM se ha utilizado el mismo método que se había usado previamente en [12], que consiste en definir los límites superior e inferior máximos posibles para cada variable. Los valores que se encuentren fuera de dichos límites se marcarán como 'NaN', y no se tendrán en cuenta para el entrenamiento y la validación del modelo. Debido a los altos porcentajes de datos perdidos comentados en el apartado anterior, los valores de la variable 'EtCO2' y de los datos de laboratorio no serán usados para labores de predicción, por lo que no se han calculado sus límites. Tampoco es necesario obtener los límites de los datos demográficos, ya que son constantes. Los límites considerados para las variables restantes se muestran en la tabla siguiente:

Tabla 17. Límites para los valores de cada variable. Fuente: [12].

	HR	**SatO2**	**Temp**	**DBP**	**MAP**	**SBP**	**Resp**
mín	40	80	30	50	40	30	8
máx	350	100	42	210	180	120	35

Capítulo 5. Método de monitorización diseñado

A la hora de contextualizar el método de monitorización diseñado, se ha tenido muy en cuenta la gran cantidad de datos perdidos. Una de las principales consecuencias de estas pérdidas es que, para muchos pacientes, hay largos periodos de tiempo en los que faltan datos para al menos una de las variables de interés. Esto supone un desafío científico en la medida en que el diseño que se haga ha de contemplar la eventualidad de que haya o no datos sin distinguir el paciente, pues en situaciones reales no se sabe qué va a pasar con los datos. En definitiva, es importante que el diseño propuesto sea versátil, esto es, capaz de adaptarse con facilidad y rapidez a diversas situaciones que se puedan dar con los pacientes.

Una buena forma de alcanzar lo planteado en el párrafo anterior es la que sigue:

1. Utilizar clasificadores base a partir de cada variable. El clasificador elegido para el método de monitorización está basado en el clasificador Bayesiano, regresor lineal, regresor logístico o SVM. Sin pérdida de generalidad, se asume un clasificador bayesiano. Una de las ventajas de esta elección es que funciona bien en la presencia de datos perdidos y *outliers,* lo que facilita su implementación frente a otras técnicas que requieren un tratamiento más exhaustivo de los datos para mantener una alta versatilidad y eficiencia en estas condiciones. Es necesario aclarar que esta forma de usar los clasificadores base permite aplicarlos en un instante de observación de variables dado y en función de si hay datos presentes o no. Es interesante observar que el resultado se asemeja a lo que se denomina clasificador Naïve Bayes, donde se asume independencia de las variables, aunque esto no sea cierto o no se haya verificado, y se combinan todos los clasificadores aplicando la independencia asumida. En este caso no se asume independencia, sino que la eventualidad de que haya datos perdidos impone y a la vez facilita el uso de los clasificadores base según proceda.
2. Lo planteado en el punto 1 deja abierta la necesidad de combinar los clasificadores útiles en un instante dado. A tal efecto, se han estudiado diversas técnicas de combinación (apartado 3.5) de clasificadores base al algoritmo simplemente usando las probabilidades de Bayes calculadas. Tal como se verá más adelante, esto mejora la calidad global del clasificador integrado en el sistema de monitorización.

El método propuesto está dividido en etapas. En la fase de diseño, la estructura del método desarrollado es la que se muestra en la Figura 21. Primero, se realiza la selección de pacientes útiles y variables, y luego se dividen los pacientes seleccionados en dos subconjuntos de datos para entrenamiento y test. Posteriormente, se ejecuta el módulo de aprendizaje, mediante el cual se generan los modelos que se van a usar en la clasificación. El método de monitorización comprende dos módulos:

1. Módulo de identificación temprana de pacientes potencialmente sépticos.
2. Módulo de detección temprana de eventos de sepsis en pacientes que tengan la enfermedad.

Los test se ejecutan sobre estos dos módulos, y finalmente se computan las métricas de calidad a partir de las etiquetas predichas.

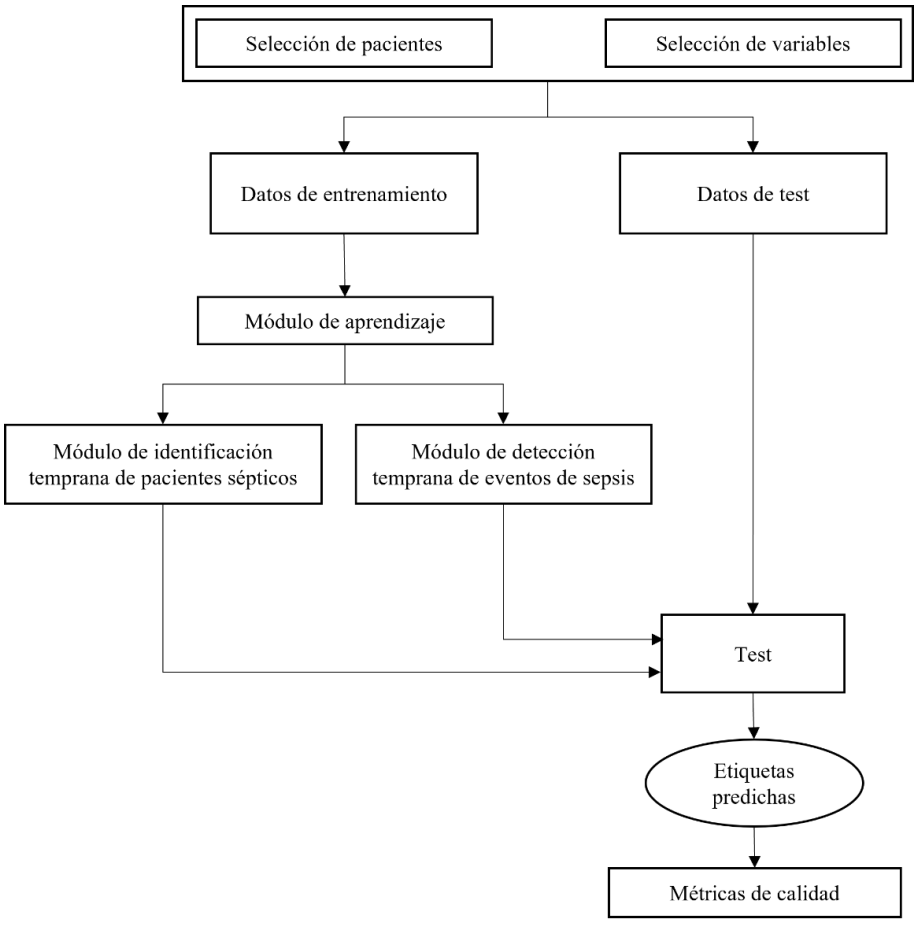

Figura 21. Diagrama de flujo del sistema de monitorización en fase de diseño.

Por otro lado, cuando el método de monitorización se esté ejecutando en un entorno en tiempo real, su estructura será previsiblemente la que se muestra en la Figura 22. En este caso, el módulo de identificación temprana de pacientes con sepsis se ejecutará primero,

y dependiendo de la detección que realice se tomarán unas acciones u otras. Si se determina que el paciente tiene la enfermedad se procederá a ejecutar el módulo de detección de eventos de sepsis, generando así las etiquetas de detección correspondientes. En caso contrario, es decir, si el paciente no se detecta como séptico, simplemente se informará a los médicos de este resultado. Dado que el método propuesto es de ayuda al médico, éste tendría la opción de mantener siempre activo el módulo de detección, aunque el primer módulo no indique que el paciente es potencialmente séptico. Tal como se verá en el apartado 5.6, esta opción se ve reforzada por una métrica de calidad dada por la función de utilidad que permite representar hora a hora cuán buena es la metodología con pacientes sin y con sepsis.

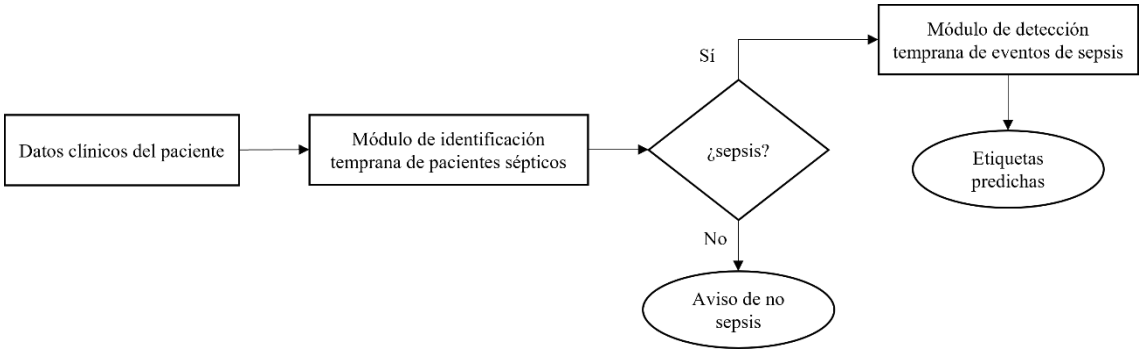

Figura 22. Diagrama de flujo del método de monitorización en tiempo real.

A continuación, se va a describir el funcionamiento de cada uno de los módulos y sistemas enumerados en esta introducción.

5.1. Selección de pacientes

En este elemento del primer módulo del sistema se realiza la selección de pacientes. Durante dicha selección, se aplican diferentes criterios para descartar aquellos pacientes que no son útiles para el diseño. De esta forma, se consiguen condiciones realistas, a la vez que se mejora la calidad de los datos usados y aumenta la eficiencia del sistema en general.

Una de las primeras cuestiones que se pueden plantear es cuál es el intervalo de tiempo en el que se va a detectar la sepsis. En este aspecto, se sabe que la función de utilidad del reto recompensa a los algoritmos capaces de predecir la sepsis entre 12 y 3 horas antes del diagnóstico clínico. También se había comentado anteriormente la eficacia de la terapia temprana dirigida por objetivos, un protocolo que se aplica durante las seis

primeras horas de sospecha de la enfermedad. A partir de esto, se puede concluir que el momento óptimo para predecir la sepsis se encuentra entre 6 y 12 horas antes de que se manifiesten los primeros síntomas. Con este fin, se ha definido el parámetro de selección "h_min_risk", que representa el número mínimo de horas que debe haber entre el ingreso de un paciente en la UCI y el diagnóstico oficial de sepsis.

A su vez, lo expuesto en el párrafo anterior hace que sea necesario considerar un número mínimo de horas en la UCI por paciente. Por ejemplo, si un paciente ha estado ingresado menos de 12 horas en la UCI, es lógico pensar que, o bien no cumple "h_min_risk", o bien cuenta con un número muy reducido de horas sépticas. A esto se suma el hecho de que, para poder realizar una predicción razonablemente buena, es necesario disponer de un histórico de datos con una extensión suficiente. Por tanto, se debe definir una cantidad adecuada de horas mediante el parámetro de selección "hora_min" que permita conservar la mayoría de los pacientes sin comprometer la utilidad de los datos.

Otro aspecto a tener en cuenta a la hora de considerar pacientes es si tienen una cantidad razonable de datos útiles. Si un paciente presenta un porcentaje demasiado alto de valores perdidos puede afectar de forma negativa al aprendizaje del modelo. Esto se comprueba mediante el parámetro de selección "per_ref", es decir, el porcentaje mínimo de datos útiles que debe tener un paciente para considerarse válido en el contexto del modelo que se va a generar y testear.

Por último, en el apartado 4.2 se comentó que las clases estaban desbalanceadas y que la disponibilidad de datos de pacientes sin sepsis es mucho mayor que en el caso de los pacientes sépticos. El desbalanceo entre dos clases es uno de los aspectos que perjudican el desempeño de un modelo de inteligencia artificial, debido a que el algoritmo puede tener dificultad para aprender los conceptos relacionados con la clase minoritaria. A estos efectos, la solución más simple y efectiva consiste en forzar la igualdad entre clases estableciendo un límite superior, que se corresponde con el número de casos que tiene la clase minoritaria. Esto se puede hacer porque se dispone de un número muy alto de pacientes de ambas clases, y lo mismo ocurre entre los pacientes sépticos donde hay un número muy alto de horas con y sin sepsis. Por ejemplo, si se tienen 1000 pacientes sépticos y 1500 pacientes no sépticos, se tomarán 1000 pacientes de cada caso para igualar la prevalencia de las clases.

En cuanto a la distribución de datos para entrenamiento y validación, se ha optado por utilizar el 20% de los pacientes para la realización de los tests, y el 80% para el entrenamiento. Siguiendo con el ejemplo anterior, se asignarían 200 pacientes sépticos y 200 pacientes sin sepsis para el conjunto de test, mientras que el conjunto de entrenamiento estaría formado por los 1600 pacientes restantes. Aunque la estrategia comentada se asemeja a lo descrito sobre el submuestreo "*single hold-out"* (apartado 3.4), no ha sido ésta la técnica de validación cruzada que se ha empleado. En su lugar, se han empleado dos variaciones de *"k-fold"*. En el apartado 5.6 se describe con mayor profundidad la estrategia de validación empleada.

5.2. Selección de variables

En este elemento del primer módulo del sistema se realiza la selección de variables. De acuerdo con lo que se comentó en el capítulo anterior, hay una gran cantidad de variables clínicas que presentan un alto porcentaje de valores perdidos (Figura 18 y 19). Concretamente, los valores de laboratorio tienen más del 80% de datos perdidos para ambos conjuntos de datos. Asimismo, 'EtCO2' tiene un 100% de datos perdidos en el conjunto de datos A, lo que dificultaría la implementación de una validación cruzada entre los dos conjuntos. Por otro lado, los parámetros demográficos permiten caracterizar al paciente, pero no aportan información sobre la evolución de la enfermedad. Atendiendo a estos criterios se ha tomado la decisión de excluir del modelo de aprendizaje las variables clínicas mencionadas y evaluar únicamente los signos vitales (Tabla 13) para la predicción de la sepsis: 'HR', 'O2Sat', 'Temp', 'SBP', 'MAP', 'DBN' y 'Resp'. Hay que destacar que, más allá del porcentaje de datos útiles que presentan, estas variables son potencialmente útiles pues es razonable pensar que tienen características específicas asociadas a la sepsis. Se ha incluido también la variable 'ICULOS' (Tabla 14), ya que, de acuerdo con lo observado en las Figura 12 y 13, los pacientes con sepsis generalmente permanecen ingresados en la UCI más tiempo que los pacientes no sépticos.

Además de estas variables que se proporcionan de forma directa en el conjunto de datos, se han calculado y añadido dos nuevas variables que se consideran de gran importancia. En el Capítulo 2 ya se comentaba que, en el contexto actual de la definición Sepsis-3, las escalas SOFA y qSOFA son los métodos principales para la predicción de sepsis. En este caso, no es posible aplicar la escala qSOFA debido a que solo se tienen dos de las tres variables necesarias. Sin embargo, es posible computar la puntuación SOFA para cada hora a partir de las variables clínicas disponibles. Esta es la primera

variable adicional que se ha añadido al modelo. Por otro lado, en [83] y [36] se destaca la importancia de la característica 'ShockIndex', definida como la frecuencia cardiaca dividida por la presión sanguínea sistólica. Ya que se dispone de los datos necesarios para su cálculo, esta variable también se ha incorporado al algoritmo desarrollado.

Con la finalidad de hacer un primer estudio acerca de si las variables clínicas seleccionadas son potencialmente útiles para diferenciar entre ambas clases de pacientes, se ha realizado un estudio inicial sobre la capacidad discriminante de las mismas. A tal efecto se ha aplicado un análisis estadístico basado en diagramas de cajas. Mediante este tipo de diagramas se tiene una forma estandarizada de mostrar la distribución de los datos basada en un resumen de cinco números ("mínimo", primer cuartil (Q1), mediana, tercer cuartil (Q3) y "máximo"). De esta forma, se obtiene información acerca de sus valores atípicos y cuáles son sus valores. También puede indicar si los datos son simétricos, su grado de agrupación, y si están sesgados y cómo. Si, además, se usa para comparar la distribución de datos de dos grupos diferentes, se puede obtener información acerca de si hay solape y si hay posibilidad de discriminar entre ambos grupos. En este caso, los grupos son pacientes sépticos ('Pac1') y pacientes no sépticos ('Pac0'). Como se puede observar en la Figura 23, en la mayoría de los casos hay una diferencia notable entre los valores de las variables para 'Pac1' y 'Pac0' en el conjunto de datos A. Esto permite adelantar que, aplicando dichas variables, es posible determinar a qué clase pertenece un paciente cualquiera. Esta situación se da también para el conjunto de datos B, como se muestra en la Figura 24. No obstante, hay variables potencialmente buenas y otras que son peores:

- La variable 'HR' presenta valores generalmente más altos para pacientes no sépticos que para pacientes con sepsis en el conjunto A, mientras que en el caso del conjunto de datos B se da la situación inversa. Es decir, 'HR' tiene poder discriminante en los conjuntos individuales, pero puede que el rendimiento del sistema sea menor si se mezclan ambos conjuntos de datos.
- La variable 'O_2' se solapa para las dos clases de pacientes, aunque se puede apreciar que los pacientes no sépticos presentan valores ligeramente superiores y más concentrados en ambos conjuntos de datos. En general, el poder discriminante individual de esta variable clínica no es muy alto.
- En el conjunto de datos A, se puede afirmar que la variable 'Tm' no permite diferenciar entre los pacientes, ya que el rango de valores que presenta para ambos

casos se solapa. De hecho, incluso la mediana está a la misma altura para los pacientes sépticos y no sépticos, por lo que el poder de clasificación de esta variable es muy pequeño. Por otro lado, en el conjunto de datos B si hay una clara diferencia entre ambas clases, por lo que la capacidad discriminante de 'Tm' en este conjunto de datos es bastante elevada.

- La variable 'SB' es otro caso donde los valores para ambas clases de pacientes se solapan en cierta medida en el conjunto de datos A. Aunque no es una situación tan extrema como la que presenta 'Tm', en términos de clasificación individual no es óptima. Para el conjunto de datos B, este solape es mucho menor, es decir, 'SB' sí tiene una capacidad discriminante moderada cuando se trata el segundo conjunto de datos.
- En ambos conjuntos de datos, la variable 'MA' presenta valores más altos para pacientes no sépticos que en aquellos que tienen la enfermedad. Concretamente, en el conjunto de datos A el solape entre clases es muy pequeño y el poder discriminante es bueno, mientras que en el conjunto de datos B hay un poco más de solape debido a que el rango de valores de los pacientes sin sepsis es más pequeño. Aun así, 'MA' tiene una capacidad discriminante potencialmente buena.
- Como en el caso anterior, la variable 'DB' presenta valores más altos para pacientes sin sepsis en ambos conjuntos de datos. Además, el solape en ambos casos es casi inexistente. Por ello, el poder discriminante de 'DB' es potencialmente alto, tanto cuando se trata de los conjuntos de datos individuales como cuando se combinan ambos.
- La variable 'Rs', por su lado, tiene valores más altos cuando se trata de pacientes sépticos. En el conjunto de datos B no existe solape entre ambas clases, mientras que en el conjunto de datos A hay una pequeña intersección en los rangos de valores. Por tanto, la capacidad discriminante de 'Rs' es potencialmente alta.
- Las variables 'IC', 'pSF' e 'InS' no parecen tener poder discriminante. Estas tres variables presentan un rango de valores muy pequeño en ambos conjuntos de datos que además se solapan entre los pacientes sépticos y los no sépticos.

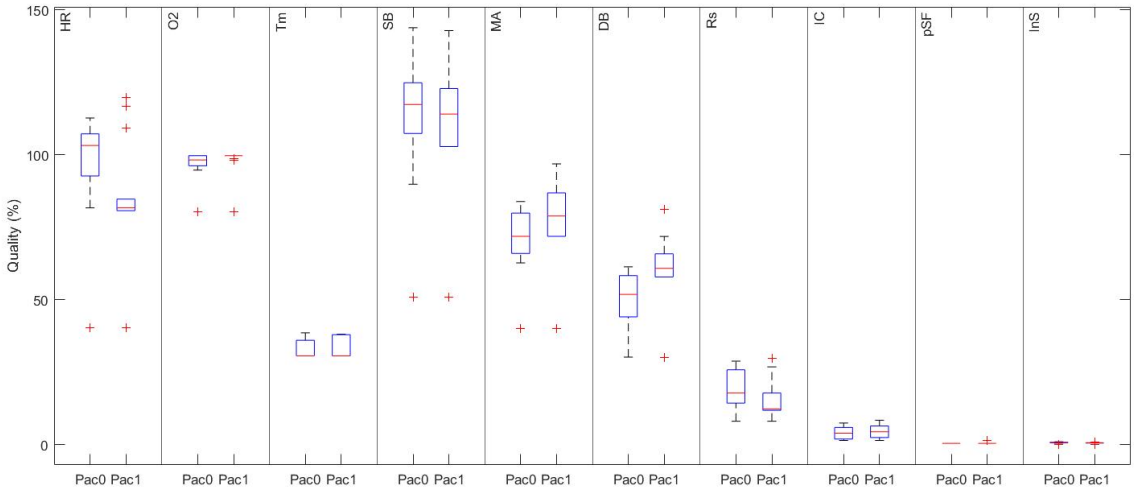

Figura 23. Diagramas de cajas de las variables seleccionadas para pacientes sépticos y no sépticos del conjunto de datos A.

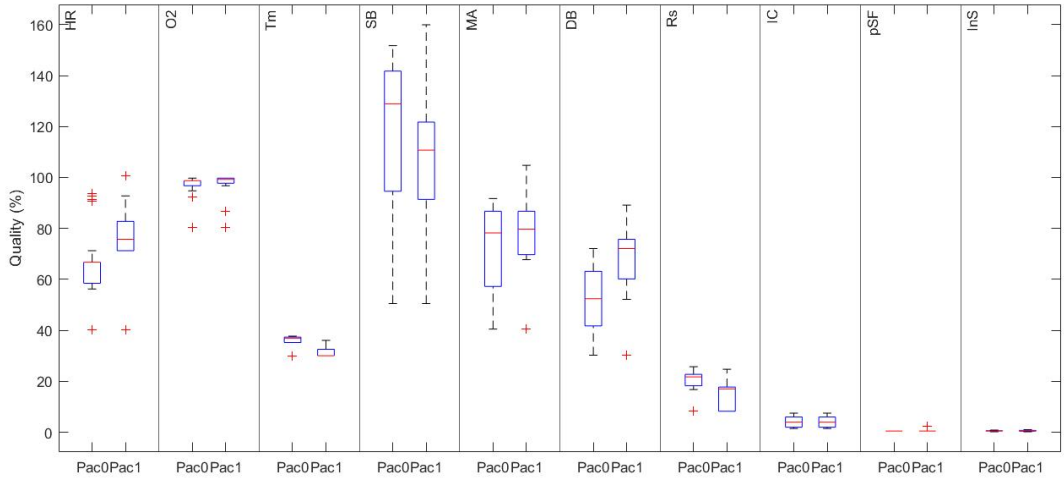

Figura 24. Diagramas de cajas de las variables seleccionadas para pacientes sépticos y no sépticos del conjunto de datos B.

El objetivo de este planteamiento es determinar si es razonable realizar un clasificador individual con cada una de estas variables clínicas. La respuesta es que sí, incluso cuando hay solape entre los valores de una variable, aunque sea esperable que el rendimiento del clasificador no sea demasiado bueno en esos casos.

Otro procedimiento que permite verificar el poder discriminante de las variables clínicas es el contraste de hipótesis. Un contraste de hipótesis es un conjunto de reglas usadas para tomar una decisión acerca de la validez de una hipótesis en base a una probabilidad. La hipótesis sobre la que se decide debe ser una hipótesis nula (H_0), es

decir, una hipótesis que se rechaza o se trata de refutar [84]. La contraparte de la hipótesis nula es la hipótesis alternativa (H_1), que es la hipótesis que se cree verdadera o que se quiere demostrar. En el contexto de este trabajo, la hipótesis H_0 sería la afirmación de que las variables clínicas anteriores no sirven para diferenciar entre pacientes sépticos y no sépticos. Por tanto, la hipótesis alternativa H_1 es lo contrario, es decir, que las variables clínicas tienen poder discriminante.

En la práctica, es común que el contraste de hipótesis se lleve a cabo mediante el cálculo del valor *p* (*p-value*). En una población, el valor *p* es la probabilidad de obtener, por azar, una diferencia tan grande o mayor de la observada en las muestras. Generalmente, se establece un umbral a partir del cual el valor de *p* proporciona el grado de credibilidad de H_0: si *p* es muy pequeño (por ejemplo, muy inferior a 0,001), significa que H_0 es del todo imposible y por tanto puede ser descartada; si el valor de *p* se encuentra entre 0,001 y el umbral significa que existen fuertes evidencias en contra de H_0, por lo que puede descartarse o no en función del valor obtenido; por último, si *p* es muy grande (superior al umbral), no se puede descartar la hipótesis, por lo que se toma como estadísticamente cierta [27]. El valor típico de este umbral es 0,01 o 0,05, ya que son probabilidades lo suficientemente pequeñas como para tomar una decisión en base a ellas.

Existen diferentes métodos para el cálculo del valor p. Sin embargo, no se puede aplicar cualquier método a cualquier conjunto de muestras de forma indiscriminada, sino que la elección de la forma de cálculo de p depende de las características de la población concreta que se está estudiando. Por ejemplo, el test de Mann-Whitney-Wilcoxon [85] tiene tres condiciones necesarias principales para poder ser aplicado a un conjunto de muestras dado:

- Independencia de los datos.
- Los datos deben ser ordinales o bien poderse ordenar de mayor a menor.
- No es necesario asumir una distribución normal de las muestras. No obstante, ambas deben tener el mismo tipo de distribución.

Las variables clínicas con las que se ha trabajado cumplen estos requisitos, por lo que el test de Mann-Whitney-Wilcoxon puede aplicarse para el cómputo del valor p de las mismas. Los resultados del valor p para ambos conjuntos de datos se muestran en las y Tabla 19. Tomando como referencia un umbral 0,01, se observa que el valor p de todas

las variables clínicas estudiadas está por debajo de éste, siendo o bien 0 o bien mucho menor que 0,01.

Tabla 18. Valor p de las variables clínicas en el conjunto de datos A.

	HR	**O_2**	**Tm**	**SB**	**MA**	**DB**	**Rs**	**IC**	**pSF**	**InS**
p-value	0	0	0	≪0,01	0	0	≪0,01	0	0	≪0,01

Tabla 19. Valor p de las variables clínicas en el conjunto de datos B.

	HR	**O_2**	**Tm**	**SB**	**MA**	**DB**	**Rs**	**IC**	**pSF**	**InS**
p-value	≪0,01	0	≪0,01	0	0	0	0	0	0	≪0,01

5.3. Módulo de aprendizaje

En el módulo de aprendizaje se realiza el entrenamiento del algoritmo diseñado para realizar una distinción entre pacientes sépticos y no sépticos, así como para monitorizar la aparición de sepsis en aquellos pacientes que se han identificado como potencialmente sépticos. Para este último fin, se ha entrenado un segundo modelo únicamente con datos de pacientes con sepsis. En esta ocasión, en lugar de diferenciar entre dos clases de pacientes, se diferencia entre horas normales y horas con sepsis de un paciente que tenga la enfermedad. La metodología de obtención de los modelos que se explica a continuación se aplica a ambos modelos.

Tomando como referencia el clasificador de Bayes, el cómputo las distribuciones de probabilidad condicionada se ha realizado a través del cálculo de frecuencias relativas, es decir, la probabilidad condicional buscada viene dada por el cociente entre el número de casos favorables y el número de casos posibles:

$$P(X_n|Y) = \frac{N_{Nn}}{N_N} \tag{22}$$

donde N_{Nn} es el número de veces que la característica n aparece en la clase Y y N_N es el número de observaciones de dicha clase. Esta estimación se ha realizado mediante una discretización de los valores de cada variable en N contenedores (N_{bins}). El número de contenedores depende del rango de valores que pueden tomar las variables en cada caso. Con este fin, las variables elegidas se transforman mediante una función de cuantificación propia desarrollada en [12], que utiliza el algoritmo de Lloyd [86] para asignar un nivel

de cuantificación a cada valor. El algoritmo de Lloyd es un método de cuantificación de señales que permite obtener el cuantificador escalar óptimo, desarrollado por Stuart Lloyd en 1957 como una técnica para la modulación por pulsos codificados (*Pulse Code Modulation*, PCM). En esencia, es un método de agrupamiento de datos en contenedores o *clusters* que clasifica elementos iterativamente en base a sus propiedades. Al inicio de la ejecución se fija un número *N* de contenedores y se asignan elementos empleando funciones como la distancia euclídea, por ejemplo. Para el cálculo de los centroides del *cluster* se emplea la media estadística.

Uno de los inconvenientes que se pueden presentar al calcular las probabilidades mediante la técnica frecuentista es que en ausencia de un valor en el conjunto de entrenamiento la probabilidad asociada será cero, independientemente de que este valor pueda estar presente en el conjunto de datos de pruebas. Esto podría provocar que las predicciones sean dudosas o incluso erróneas. Para evitar o mitigar este problema, una de las técnicas más usadas es el suavizado de Laplace [87]. En este caso, la probabilidad condicional $P(X_n|Y)$ viene dada por el número de casos favorables más uno dividido por el número de observaciones más el número de contenedores o centroides, N_{bins}. Así, la probabilidad condicional queda de la siguiente forma:

$$P(X_n|Y) = \frac{N_{Nn} + 1}{N_N + N_{bins}} \quad (3)$$

La cuantificación y el suavizado de Laplace se aplica por separado a pacientes sépticos y no sépticos. Esto aporta flexibilidad en la medida de que permite realizar la clasificación mediante una comparación de probabilidades aplicando cocientes, logaritmos, etc. Para el cálculo del umbral de decisión óptimo, se han obtenido las curvas ROC (apartado 3.8.1), usando como parámetro ajustable las “h_min_risk” primeras horas de los pacientes del conjunto de entrenamiento. El umbral calculado se emplea en el módulo de monitorización de pacientes potencialmente sépticos, que se describe en el apartado siguiente.

5.4. Identificación temprana de pacientes potencialmente sépticos

La primera utilidad del programa diseñado es la monitorización de pacientes. Dicha monitorización consiste en un análisis hora a hora de los pacientes hasta detectar un

posible caso de sepsis. Esto se realiza mediante dos pasos principales: cálculo de la probabilidad acumulada y comparación del resultado con el umbral de decisión óptimo. El diagrama de flujo mostrado en la Figura 26 representa el funcionamiento de este módulo. A grandes rasgos, el programa realiza dos tareas principales:

- En primer lugar, cada hora se obtiene, para cada valor de las variables clínicas, un par de probabilidades asociadas que da lugar a una clasificación.
- En segundo lugar, se determina cuál es la clase en la que se va a clasificar el paciente mediante la comparación de las probabilidades anteriores.

Como se había mencionado anteriormente, una estrategia de combinación adecuada favorece la precisión del algoritmo. Se habían considerado tres estrategias principales: combinación lineal, producto y *majority voting*. No obstante, en la práctica se ha optado por usar dos estrategias y combinar las clasificaciones obtenidas. En la Figura 25 se muestran los rangos intercuartiles de las principales métricas obtenidos para las cuatro estrategias de combinación tras repetir la clasificación un total de 100 veces. De este diagrama se puede observar que:

1. La combinación producto (B) presenta la mejor especificidad.
2. La estrategia de *majority voting* (C) ofrece una sensibilidad cercana al 100%.
3. Al combinar ambas estrategias (D), se ha obtenido una mejora sustancial de la especificidad, manteniendo una sensibilidad razonablemente buena.

Tabla 20. Unión de estrategias de combinación producto y majority voting para el módulo de monitorización de pacientes.

Producto	*Majority Voting*	Resultado
0	0	0
0	1	0
1	0	0
1	1	1

En la Tabla 20 se muestra la combinación de ambas estrategias. Como se observa, de esta forma solo se clasificará sepsis cuando los resultados de ambas estrategias de combinación coincidan.

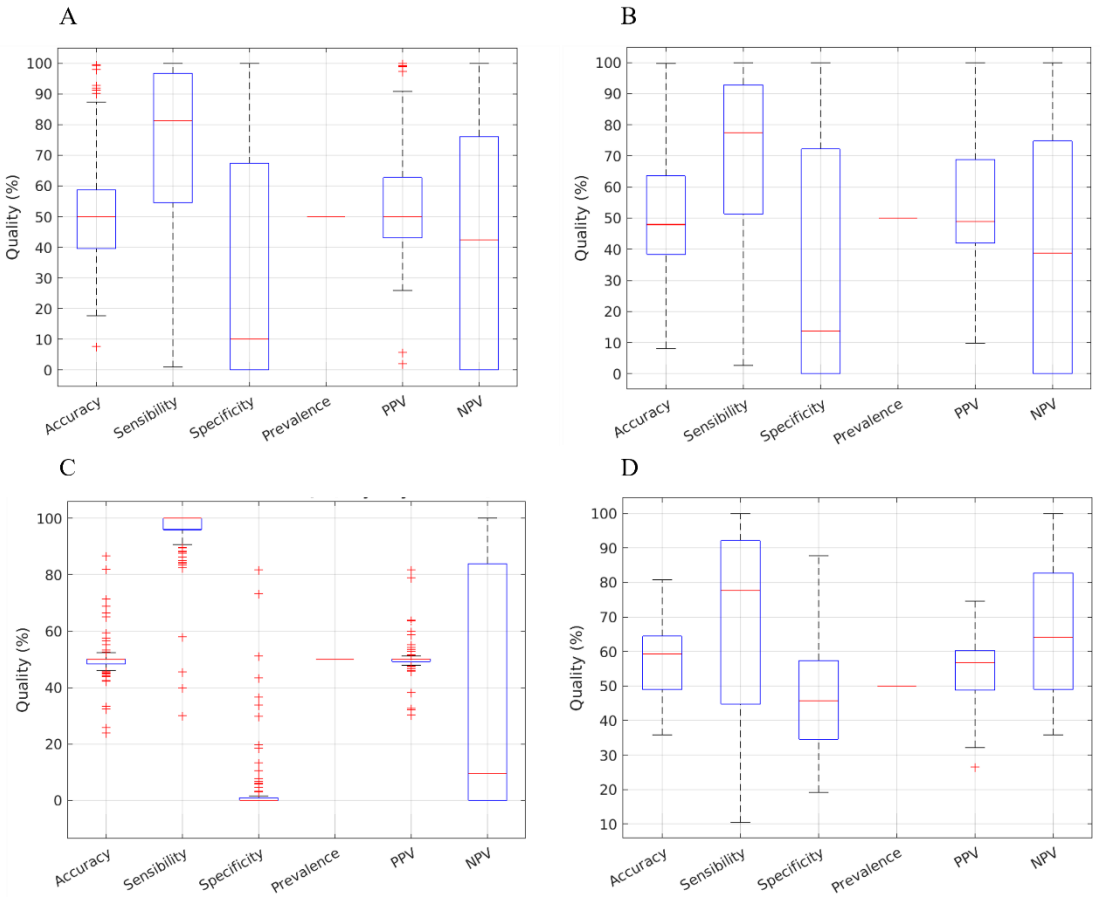

Figura 25. Diagramas de cajas de las principales métricas para las diferentes estrategias de combinación. (A: Lineal; B: Producto; C: Majority Voting; D: Híbrido).

A la vista de estos resultados, cabe preguntarse si es posible mejorarlos. La respuesta es sí, mediante un posprocesado de las probabilidades. Esto se consigue incluyendo en el sistema el concepto de probabilidad de detección acumulada. Hay que recordar que la probabilidad de detección acumulada, como se había comentado en el apartado 3.6, es la probabilidad de realizar una detección después de una serie de oportunidades. Por lo general, dicha probabilidad se calcula dentro de una ventana deslizante de tamaño fijo. De esta forma, la estructura final del módulo, tras incluir la probabilidad acumulada, sería la que se muestra en la Figura 26.

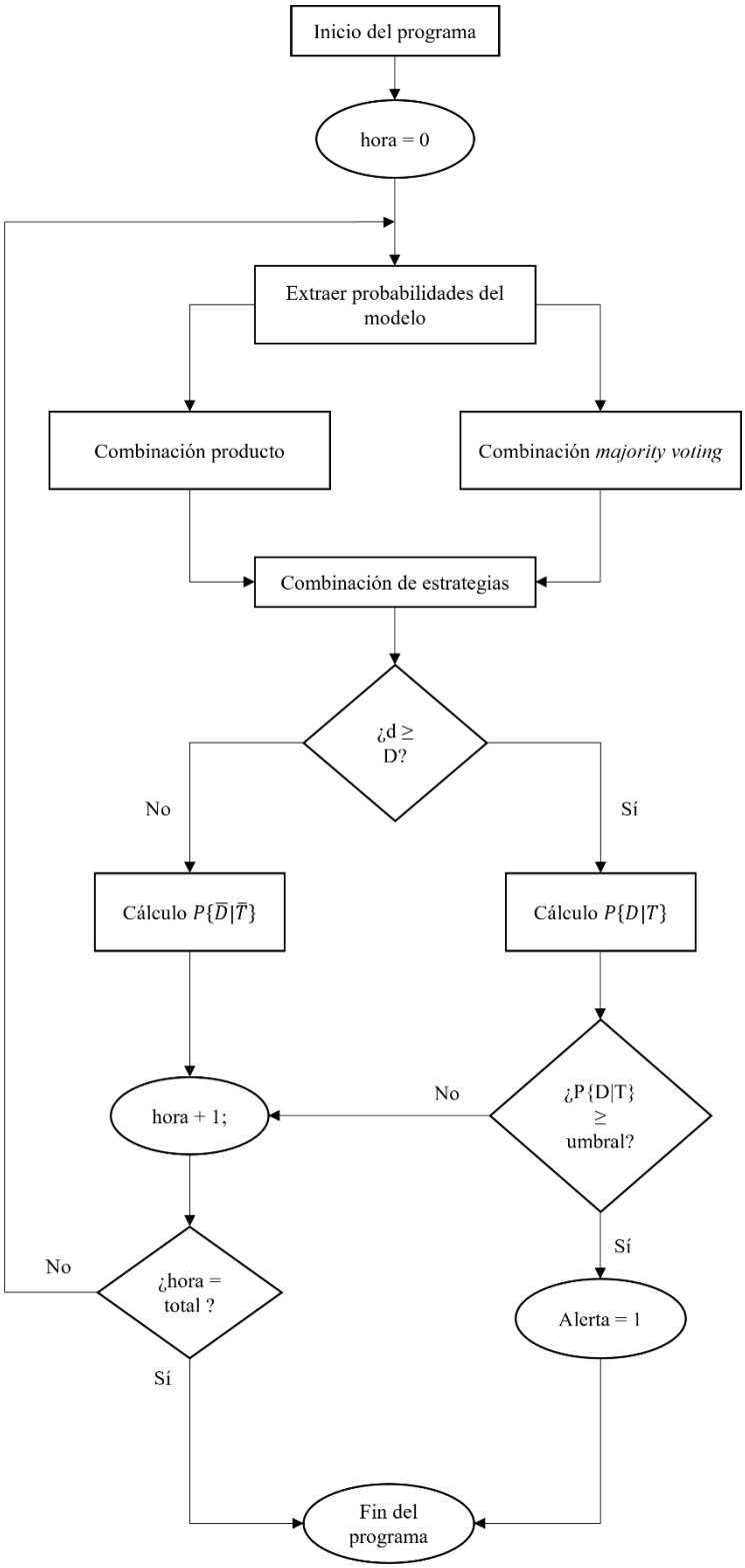

Figura 26. Diagrama de flujo del módulo de identificación temprana de pacientes potencialmente sépticos.

El funcionamiento de la probabilidad acumulada en este programa es el que sigue:

- La probabilidad acumulada se calcula en una ventana deslizante de tamaño "h_min_risk", es decir, la monitorización del paciente comienza un determinado

número de horas después de su ingreso en la UCI. Se ha optado por esta fórmula debido a que es necesario contar con un histórico de datos de longitud razonable para poder obtener un resultado creíble de la probabilidad acumulada.

- Como se había comentado en el apartado 3.6, se va a dar tanto la probabilidad de tener la enfermedad teniendo un test positivo, $P\{D|T\}$, como la probabilidad de no tenerla teniendo un test negativo, $P\{\bar{D}|\bar{T}\}$. El criterio para calcular una u otra probabilidad es realizar una serie de detecciones en la ventana analizada.
- Si el número de detecciones, *d*, es mayor o igual al mínimo establecido, *D*, se obtendrá la probabilidad $P\{D|T\}$. En caso contrario, se realizará el cómputo de $P\{\bar{D}|\bar{T}\}$.
- Cuando la probabilidad de un caso positivo supera el umbral óptimo determinado mediante las curvas ROC, se desatará una alerta, finalizando la monitorización del paciente.
- En caso de que la probabilidad acumulada no supere el umbral, el programa finalizará sin haber lanzado ninguna alerta.

En la Figura 27 se muestra un ejemplo donde se demuestra que incorporando la probabilidad de detección acumulada al algoritmo se consigue suavizar las falsas alarmas durante el período normal, así como las omisiones en el período de sepsis. Asimismo, la probabilidad de detección acumulada permite realizar detecciones más estables durante el período de detección temprana, y, por tanto, más fiables.

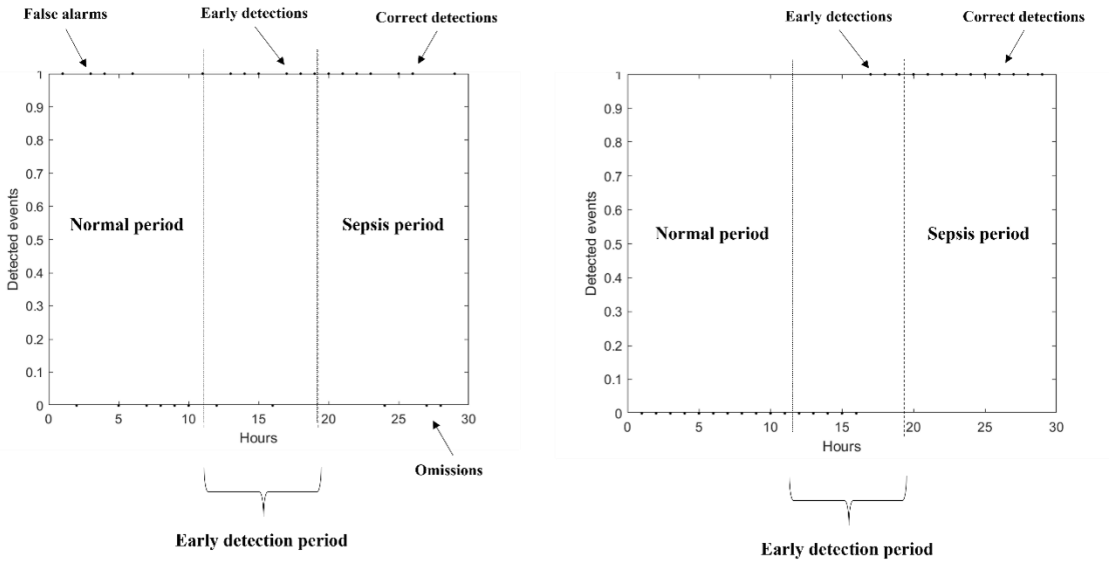

Figura 27. Eventos detectados en un paciente con sepsis (Izq: Sin usar probabilidad acumulada; Dcha: Usando probabilidad acumulada).

5.5. Módulo de detección temprana de eventos de sepsis

El módulo de identificación temprana de eventos de sepsis en pacientes con la enfermedad realiza una clasificación hora a hora de la sepsis. La estructura del programa es la que se presenta en la Figura 28. El funcionamiento del módulo es el siguiente:

- El programa se inicia cuando se realiza una detección en el módulo de identificación temprana de pacientes potencialmente sépticos. No obstante, este módulo podría activarse independientemente del resultado de identificación temprana.
- Se realiza una monitorización hora a hora del paciente identificado para detectar horas sépticas frente a horas normales. Para ello, se hace uso del modelo entrenado exclusivamente con datos de sepsis descrito en el apartado 5.3. En principio, la clasificación se basaba en un sistema híbrido análogo al que se comentó en el apartado anterior. Sin embargo, los resultados obtenidos con este enfoque eran ligeramente inferiores a *majority voting*, por lo que finalmente se ha optado por no aplicar dicha estrategia y usar puramente *majority voting*. En el Anexo I se pueden consultar las matrices de confusión obtenidas aplicando la estrategia de *majority voting* a la identificación de eventos de sepsis.
- Una vez que se han obtenido las etiquetas predichas de un paciente, se pueden obtener las métricas de calidad. En este caso, en coherencia con el reto de Physionet, se ha aplicado la métrica basada en la función de utilidad, y se ha obtenido la utilidad del algoritmo para dicho paciente. Este es el caso que se refleja en la Figura 28. No obstante, en el reto la función de utilidad se aplicó sobre las etiquetas predichas para los pacientes del experimento. Por tanto, en fase de diseño las etiquetas de cada paciente serán almacenas para el cálculo posterior de la utilidad total del experimento.

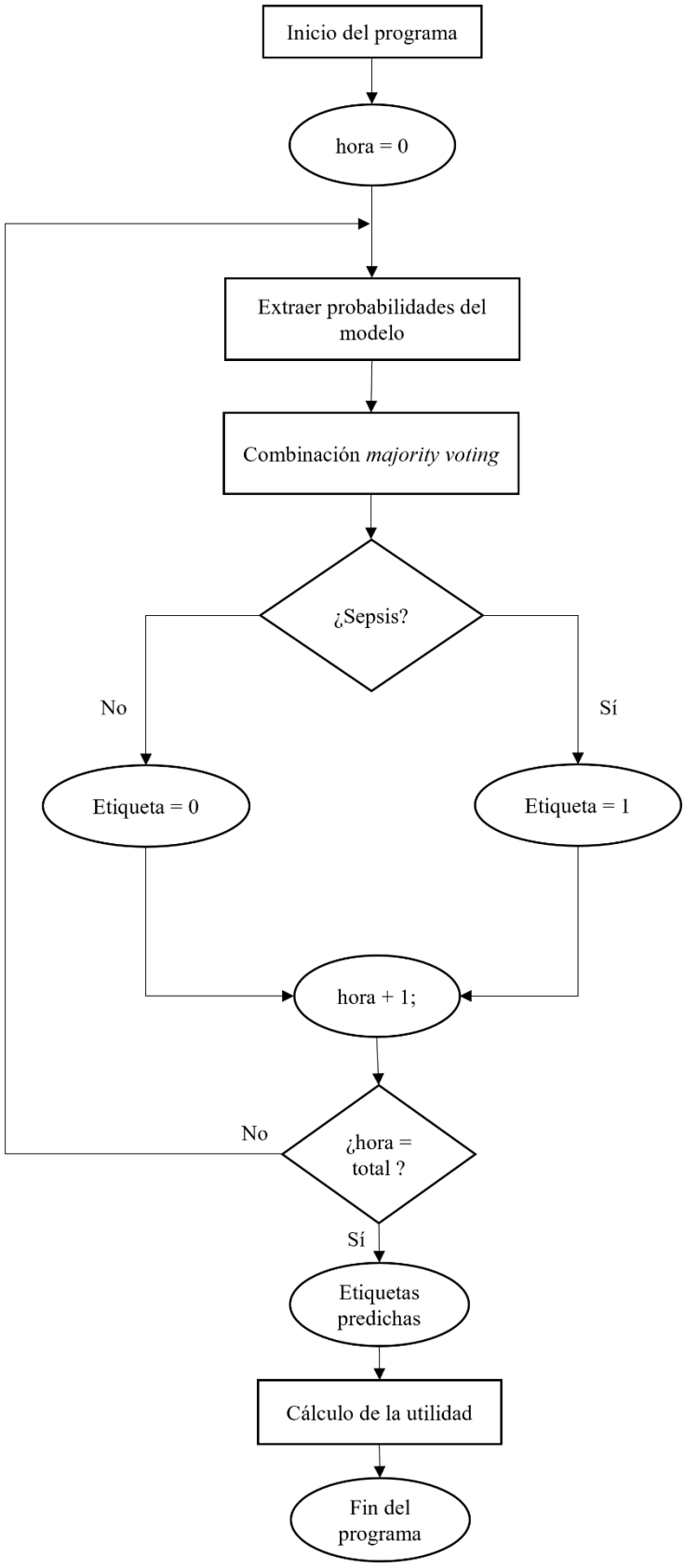

Figura 28. Diagrama de flujo del módulo de detección temprana de eventos de sepsis.

5.6. Evaluación del modelo

En el 0 se describieron los principales métodos de evaluación. En el caso del sistema diseñado, se usa principalmente el submuestreo aleatorio *k-fold*, con el que se entrena y testea con un único conjunto de datos cada vez. Concretamente, los experimentos se repiten 100 veces, y en cada iteración los pacientes seleccionados son agrupados en entrenamiento y test de forma completamente aleatoria. Hay que destacar que la distribución de pacientes es 80% para entrenamiento y 20% para test.

A lo largo de este capítulo se han mencionado diferentes parámetros que han sido usados bien para realizar la selección de pacientes o bien para mejorar la clasificación. En la se muestra una breve descripción de dichos parámetros y los valores que se han fijado para cada uno.

Tabla 21. Valores de los principales parámetros definidos para los experimentos.

Parámetro	Descripción	Valor
'K'	Número de oportunidades para realizar una detección	8
'D'	Número mínimo de detecciones necesarias	2
'h_min_risk'	Mínimo número de horas antes del diagnóstico de sepsis	8
'hora_min'	Mínimo número de horas en la UCI	24
'per_ref'	Mínimo porcentaje de datos útiles para un paciente	10

En este punto es necesario aclarar la utilidad del parámetro de selección 'hora_min' para evitar confusión. El número de horas mínimo en la UCI es un parámetro que se usa para la generación de los modelos en la fase de diseño. El hecho de que se haya establecido a un valor fijo no significa que el sistema final requiera que el paciente haya estado ingresado durante ese tiempo antes de comenzar a funcionar. El método de monitorización se ejecuta desde la primera hora y comienza a dar resultados de probabilidad a partir de la hora 'h_min_risk'. Es decir, partiendo de los valores fijados en la tabla anterior, el sistema comenzaría a realizar la clasificación de paciente séptico o no séptico ocho horas después del ingreso en la UCI de un paciente y no 24 horas después.

Una vez fijados los valores de los parámetros principales, se ha realizado la evaluación del sistema de monitorización a través de varios test. La primera cuestión que ha de verificarse es si la combinación de clasificadores aporta mejores resultados que los

clasificadores base para la clasificación de pacientes en la clase sepsis o no sepsis. Para ello, se han obtenido y comparado las principales métricas de calidad descritas en el 0 para los clasificadores base y las estrategias de combinación, en el contexto del módulo de identificación temprana de pacientes potencialmente sépticos. En las Figura 29 y 30 se muestran los rangos intercuartiles de los clasificadores base para los conjuntos de datos A y B, respectivamente. Se puede apreciar a simple vista la existencia de numerosos *outliers* (marcados en rojo) en las figuras. En este caso, los *outliers* son pacientes estadísticamente atípicos. Esto es, no están descritos de la misma forma por los modelos. Asimismo, la naturaleza alargada de las cajas representadas indica que hay una gran dispersión en los resultados de las métricas. En general, la principal dificultad que presentan los clasificadores base reside en la determinación del clasificador óptimo, ya que cada uno presenta fortalezas y debilidades en su desempeño.

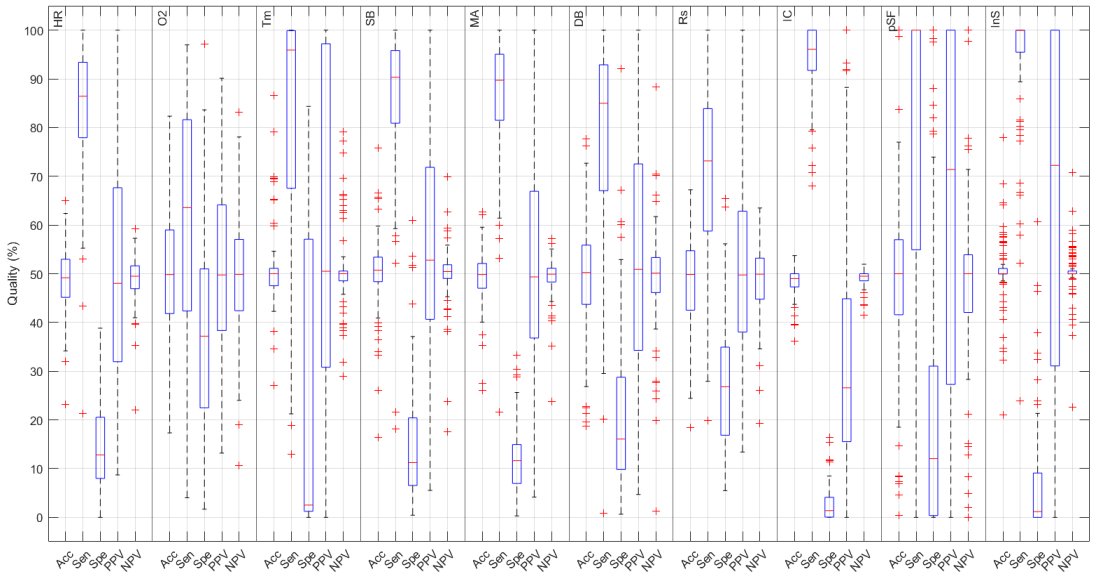

Figura 29. Diagramas de cajas de las principales métricas para los clasificadores base del conjunto de datos A.

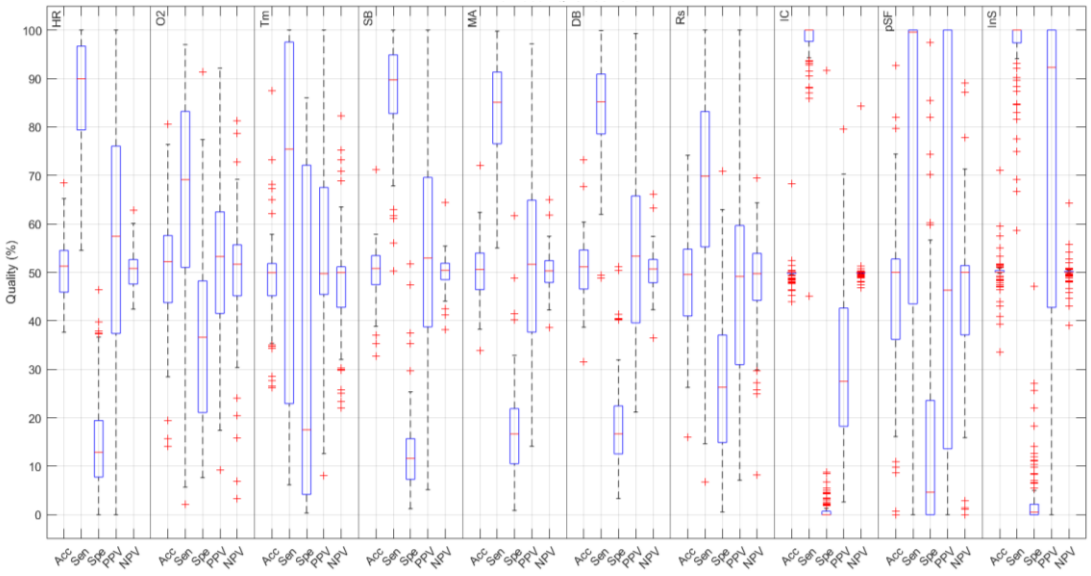

Figura 30. Diagramas de cajas de las principales métricas para los clasificadores base del conjunto de datos B.

En la Figura 31 se muestra, por otro lado, el resultado de usar las estrategias de combinación para la clasificación de los pacientes. Para ello, se han usado todos los clasificadores base comentados y se han aplicado las estrategias de combinación sobre ellos. Es necesario recordar que la combinación de clasificadores base ayuda a reducir el efecto de los datos perdidos y *outliers* en los resultados, y mejorar en general la eficiencia y fiabilidad del algoritmo. En este caso, el número de *outliers* es menor, y en general el rango de valores que contiene las métricas es más estable. Las estrategias de combinación lineal (A), producto (B) y *majority voting* (C) presentan una sensibilidad buena. No obstante, su especificidad es mejorable. El caso de *majority voting* es especialmente extremo, con una sensibilidad alrededor del 100% y una especificidad de casi el 0%. Se ha resaltado varias veces que, en el contexto médico en el que se ubica este trabajo, la gravedad de las falsas alarmas es mucho menor que la de las omisiones. Sin embargo, si la cantidad de falsas alarmas es demasiado alta, el sistema no es realmente útil. Lo interesante en un sistema de estas características es alcanzar el mejor compromiso posible entre sensibilidad y especificidad de manera que los resultados obtenidos para ambas sean lo suficientemente buenos como para afirmar que el sistema en cuestión tiene utilidad médica. Con este objetivo en mente, se ha aplicado la estrategia de combinación híbrida (D) de producto junto con *majority* voting descrita en el apartado 5.4.

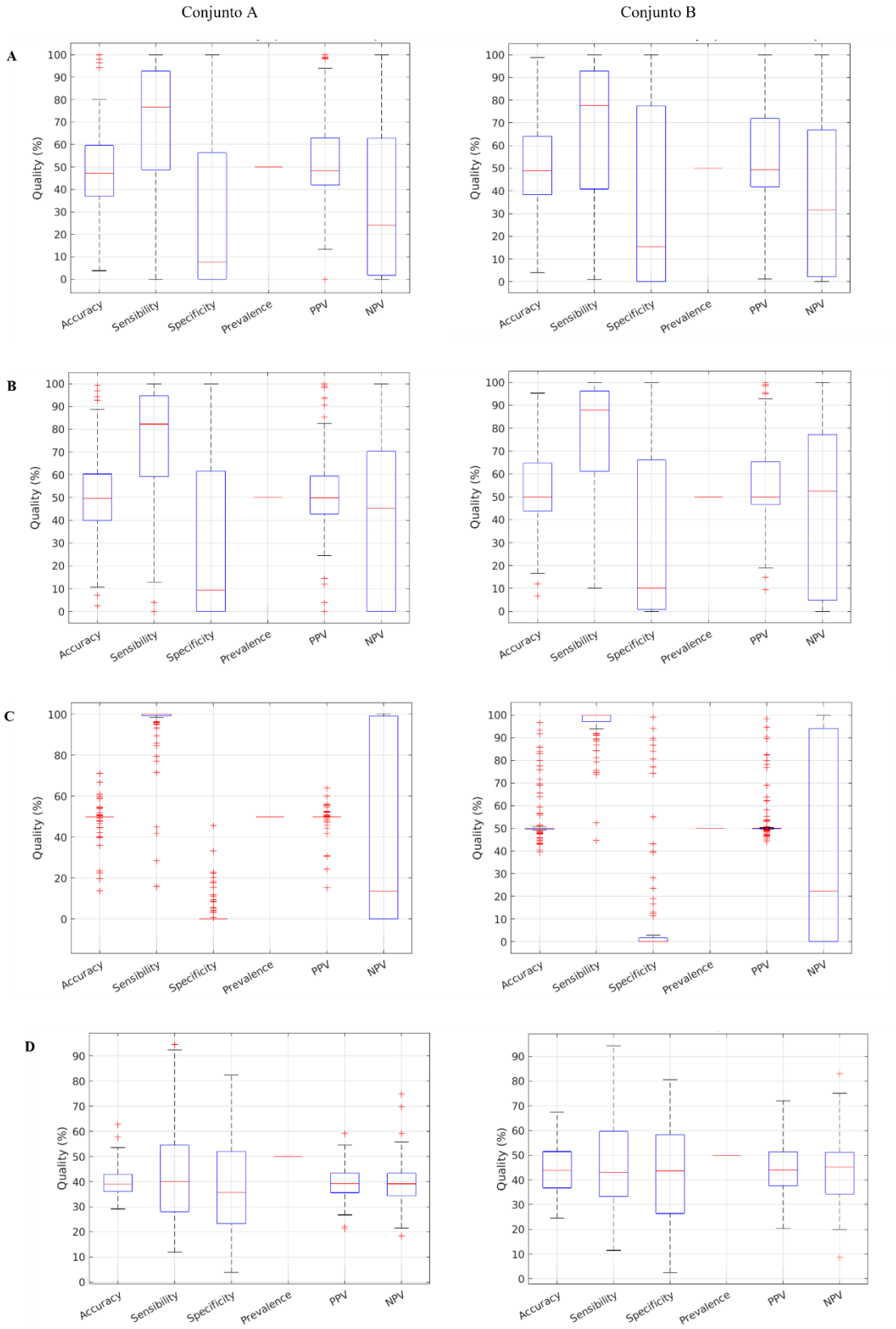

Figura 31. Diagramas de cajas de las principales métricas para las cuatro estrategias de combinación descritas. (A: Lineal; B: Producto; C: Majority Voting; D: Híbrido).

La estrategia de combinación híbrida conlleva una pérdida importante de sensibilidad respecto al resto de estrategias consideradas. Como se adelantó en el apartado 5.4, una forma de mejorar los resultados es mediante el uso de la probabilidad acumulada. Tal y como se muestra en la Figura 32, la aplicación de las probabilidades acumuladas mejora significativamente la sensibilidad respecto a lo mostrado en la Figura 31.D, manteniendo una especificidad tolerable.

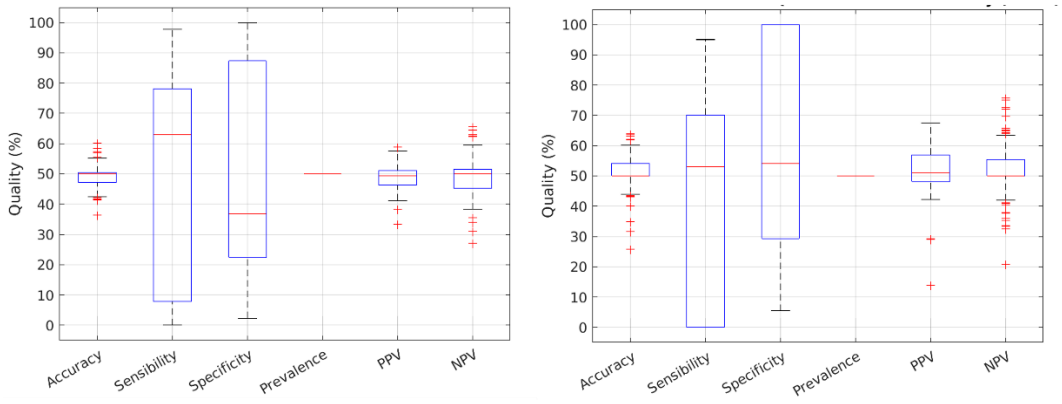

Figura 32. Diagramas de cajas de las principales métricas aplicando la probabilidad acumulada (Izq: Conjunto A; Dcha: Conjunto B).

La última métrica de calidad que se ha usado para evaluar el modelo ha sido la función de utilidad del reto de PhysioNet (apartado 3.8.6). Hay que recordar que el resultado de la función de utilidad fue el principal criterio para seleccionar a los ganadores del reto, por lo que tiene un gran peso a la hora de decidir si el sistema de monitorización diseñado es competitivo o no en el marco en que se contextualiza este trabajo. La función de utilidad se aplica a la salida del módulo de identificación de eventos de sepsis, y recibe las predicciones realizadas por el clasificador para cada paciente junto con sus etiquetas reales. Ya que cada experimento consiste en 100 iteraciones, cuando finalice el programa se tendrán 100 valores de utilidad. Esto es lo que se representa en la Figura 33. Se puede observar que la utilidad se encuentra aproximadamente entre 0,4 y 0,7 en todos los casos. Concretamente, se tiene una media de 0,603 en el caso del conjunto de datos A, 0,520 en el caso del conjunto de datos B y 0,557 en el caso en que se mezclan ambos conjuntos, A+B.

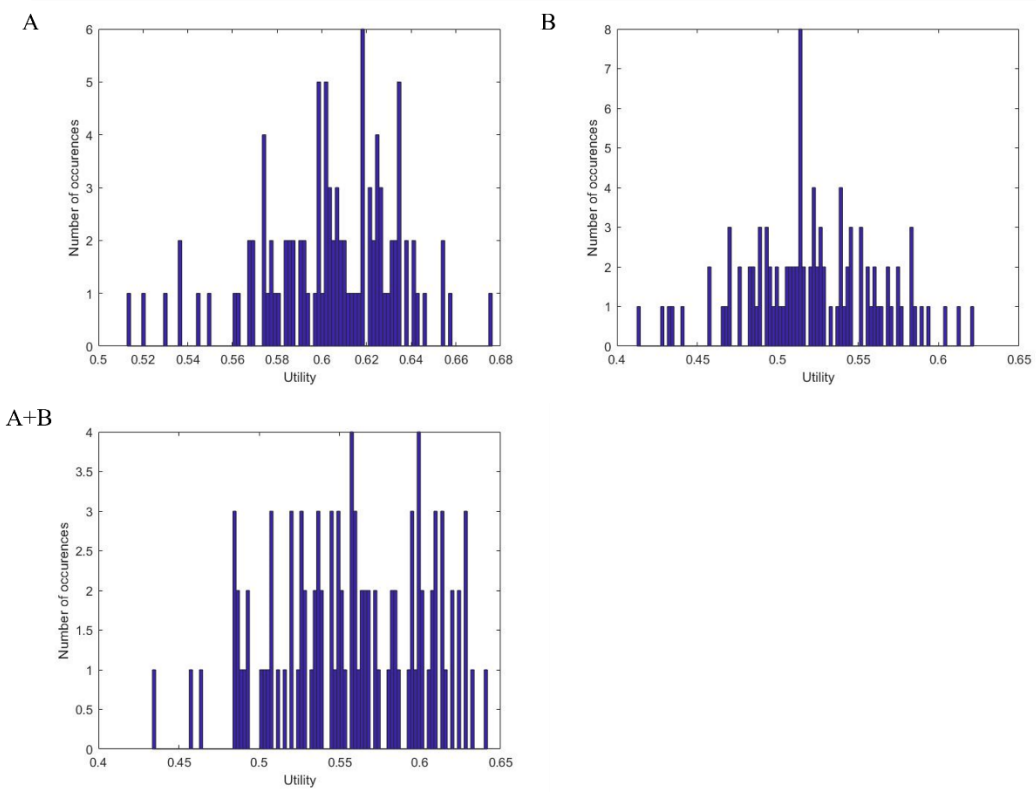

Figura 33. Utilidad total en cada conjunto de datos.

Como se ha mencionado en el párrafo anterior, esta utilidad total se calcula sobre las predicciones de todos los pacientes, es decir, se tienen en cuenta tanto los pacientes sépticos como los no sépticos. Esta fue la metodología que se aplicó para la evaluación de los participantes del reto. Sin embargo, es posible obtener una utilidad diferenciada para cada clase, como se muestra en la Figura 34. En este caso, en lugar de 100 valores de utilidad, se tiene un valor de utilidad por cada uno de los pacientes seleccionados para la fase de test en las diferentes iteraciones. Por tanto, al finalizar los experimentos se habrán obtenido varios cientos de valores de utilidad para cada una de las clases (pacientes sépticos y no sépticos). Aunque estos resultados de utilidad no se pueden comparar de forma directa con los ganadores del reto, sí permiten analizar más detalladamente la utilidad del algoritmo en los diferentes casos que se presentan. En la Figura 34 se muestran los mismos resultados de la Figura 33, pero diferenciando los pacientes de cada tipo. Esto permite estudiar la bondad de metodología propuesta con cada uno de ellos. Es interesante, por ejemplo, observar la distribución de la función de utilidad en ambos casos. La asimetría, o *skewness*, es la distribución de la variable respecto a la media aritmética. La asimetría viene dada por los coeficientes de asimetría, que indican si hay el mismo número de elementos a ambos lados de la media. Existen

diferentes coeficientes, pero este en concreto estudio se ha utilizado coeficiente de asimetría de Fisher (C_{AF}). Dependiendo del signo del coeficiente, existen tres tipos de asimetría posible [88]:

- Si $C_{AF} < 0$ la distribución tiene asimetría negativa y se concentra en valores superiores a la media.
- Si $C_{AF} = 0$ la distribución es simétrica.
- Si $C_{AF} > 0$ la distribución tiene asimetría positiva y se concentra en valores inferiores a la media.

En el caso de la función de utilidad, lo deseable es tener asimetría negativa en ambas clases, es decir, que la utilidad de los pacientes sépticos se concentre en dirección a uno y la de los pacientes no sépticos se concentren hacia cero. En la se muestra el coeficiente de asimetría calculado para los tres conjuntos de datos. En el caso de los pacientes sépticos, los tres coeficientes tienen signo negativo, es decir, la utilidad tiene asimetría negativa en estos casos. Para los pacientes que no tienen sepsis no se cumple esto, ya que solo se tiene un coeficiente negativo en el caso del conjunto de datos B, siendo el resto de los coeficientes superiores a cero.

En la también se muestra la utilidad media obtenida en los seis casos. De esta forma, la utilidad media para los pacientes sépticos supera la media de la utilidad total que se había comentado anteriormente, ya que los valores negativos de los pacientes que no tienen la enfermedad no influyen en este caso. Asimismo, se puede valorar la utilidad sobre los pacientes no sépticos, cuya utilidad óptima sería cero. Teniendo en cuenta los valores medios que se han obtenido, se puede afirmar que la utilidad del algoritmo es muy buena.

Tabla 22. Valores de las medias y skewness de la utilidad para las dos clases.

Conjunto de datos	**Pacientes no sépticos**		**Pacientes sépticos**	
	Media	*Skewness*	*Media*	*Skewness*
A	-0,031	0.4612	0,663	-1.0528
B	-0,022	-0.1414	0,556	-0.9486
A+B	-0,026	0.2281	0,604	-1.0161

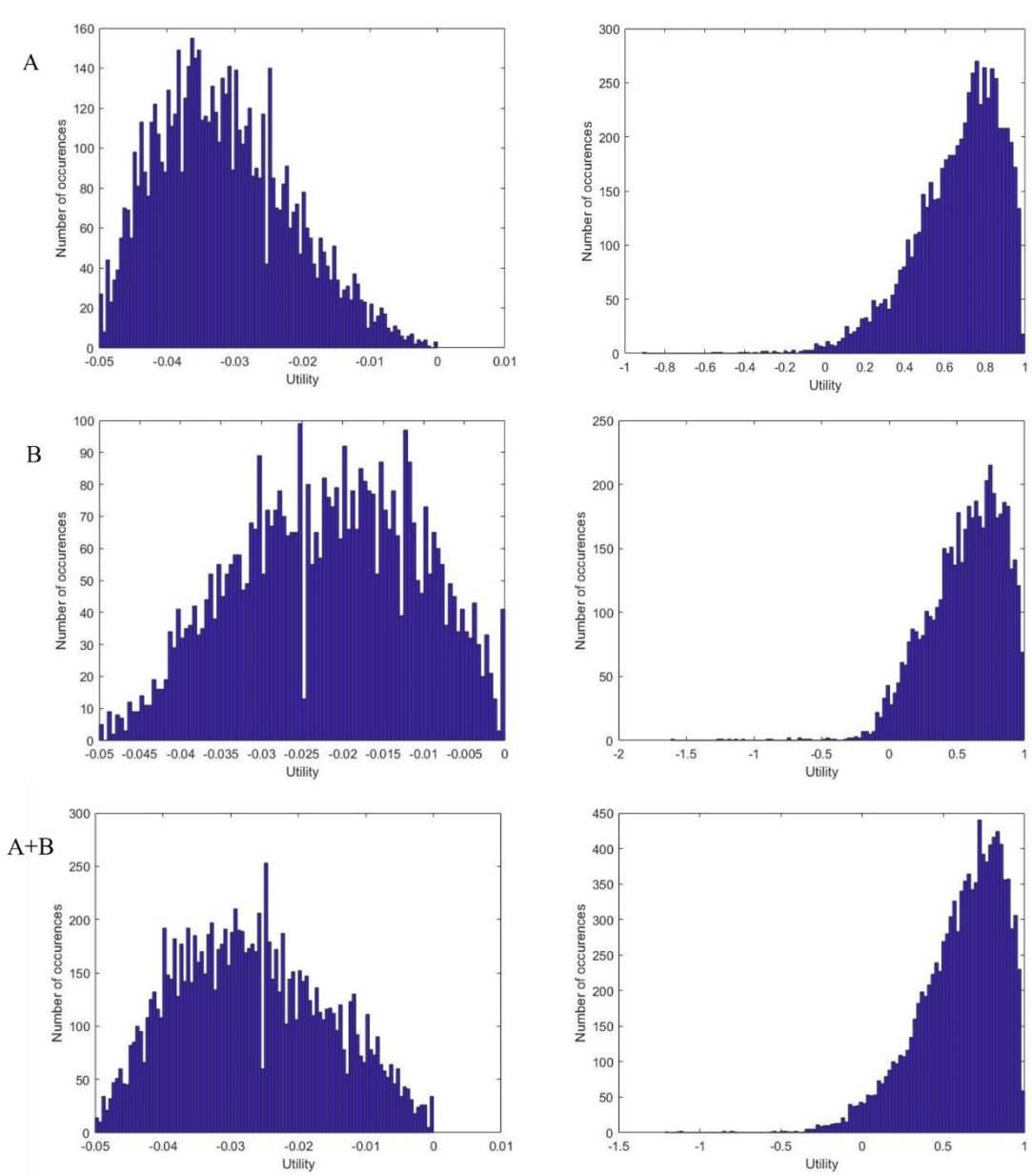

Figura 34. Utilidad por clase en cada conjunto de datos (Izq: Pacientes sin sepsis; Dcha: Pacientes sépticos).

Siguiendo con el análisis de la utilidad del algoritmo en los diferentes casos que se presentan, cabría preguntarse sobre la calidad del algoritmo de identificación temprana de la sepsis. En este caso, se hace considerando toda la estancia del paciente, y sería una forma de análisis retrospectivo.

A partir de los resultados de la función de utilidad es posible obtener un indicador de la calidad del método desarrollado adoptando un enfoque frecuentista, especialmente en el caso de los pacientes sépticos. Se ha tomado una utilidad de referencia $\mu = \mu_0 = 0$. Al cometer una predicción errónea se resta un cierto valor a la utilidad del paciente, es decir,

que una utilidad menor que 0 implicaría que se ha cometido una serie importante de errores de clasificación durante el análisis de un paciente. Así, se podría considerar que la probabilidad de error cuando el paciente es séptico es la proporción de valores de utilidad menores que 0, es decir, $P_{E1} = P(\mu \leq 0/Pac1)$.

Estos resultados son los que se expresan en la . Los resultados de probabilidad concuerdan con lo mostrado en las gráficas anteriores, ya que P_{E1} no supera el 2,86% de los casos. Dicho de otra manera, la proporción de valores de utilidad positivos es $\geqslant$ 97,14% (100-2,86 %).

Tabla 23. Proporción de valores negativos en la función de utilidad.

Conjunto de datos	$\boldsymbol{P_{E1}}$
A	0,0075
B	0,0286
A+B	0,0253

Por otro lado, en la se muestra el mayor valor de utilidad que se ha obtenido para cada conjunto de datos frente a los ganadores del reto de PhysioNet comentados en el apartado 2.4, así como la utilidad alcanzada por el trabajo previo realizado por Alba Manso [12]. También, se muestran las medias y desviaciones de la función de utilidad del método propuesto en este trabajo en cada conjunto de datos (A, B y A+B). Se puede ver que los valores de desviación son muy pequeños (un orden de magnitud menor que las medias), de lo que es posible deducir que el método propuesto es estable y fiable en los objetivos alcanzados. Por otro lado, se observa también que la utilidad conseguida, tanto la máxima como la media, en todos los casos supera con creces al ganador del reto de Physionet [36], aunque hay que destacar que, al no disponer del conjunto de datos C que se usó para validar los algoritmos durante el reto, no se puede realizar una comparación en igualdad de condiciones con los algoritmos participantes. Sin embargo, los resultados obtenidos son buenos e indican que el sistema propuesto en este TFM es muy competitivo.

A modo de ampliación, en el Anexo II se puede consultar otros valores de utilidad obtenidos para 'hora_min' igual a 48 horas. No se ha apreciado una mejoría en los resultados al realizar una selección más estricta de pacientes, por lo que se ha mantenido el valor de dicho parámetro a 24 horas.

Tabla 24. Utilidad obtenida del sistema propuesto frente a los ocho primeros clasificados del reto de PhysioNet.

Modelo empleado	Utilidad	Media	Desviación
Sistema de monitorización propuesto – Conjunto A	0,675	0,603	0,0306
Sistema de monitorización propuesto – Conjunto A+B	0,641	0,557	0,0459
Sistema de monitorización propuesto – Conjunto B	0,621	0,520	0,0415
Algoritmo para la identificación temprana de la sepsis [12]	0,522		
Modelo de regresión basado en la firma de característica [36]	0,360		
Redes neuronales	0,345		
Conjunto de modelos XGBoost [37]	0,339		
Modelo Time-Phased [38]	0,337		
Modelo basado en XGBoost [39]	0,337		
Árboles de decisión basados en refuerzo	0,332		
Modelo basado en XGBoost [40]	0,331		
Arquitectura neuronal [41]	0,328		

Por otro, en el Anexo III se presentan los valores de utilidad conseguidos utilizando la estrategia de combinación producto más *majority voting*. Como se comentó en la sección 5.6, la utilidad de los pacientes sépticos y la utilidad conjunta son ligeramente inferiores a las conseguidas aplicando únicamente *majority voting,* mientras que en el caso de los pacientes no sépticos ha mejorado. En cualquier caso, se trata de resultados superiores a los primeros clasificados del reto.

5.7. Evaluación de otros clasificadores

Los resultados mostrados hasta ahora corroboran que el método basado en el clasificador Bayesiano es muy competitivo. Ahora bien, cabe preguntarse si Bayes es realmente la mejor opción o si por el contrario es posible obtener mejores resultados usando un clasificador distinto. A estos efectos, se ha comparado la utilidad del algoritmo propuesto frente a las técnicas comentadas en el apartado 3.7: regresión lineal, regresión

logística y SVM. Para cada una de estas técnicas se han combinado los resultados de los clasificadores base usando *majority voting* y se ha aplicado la función de utilidad a las etiquetas predichas. Este proceso se ha repetido 100 veces para cada experimento. Los resultados obtenidos han demostrado que el rendimiento de Naïve Bayes es muy superior en comparación al resto de clasificadores analizados.

5.7.1. Regresión lineal

La regresión lineal es una técnica que intenta predecir el comportamiento de un conjunto de variables. Como ya se comentó en el apartado 3.7.1, se puede decir que existe regresión de los valores de una variable con los de otra cuando hay alguna línea de regresión que se ajusta en mayor o menor medida a los valores observados.

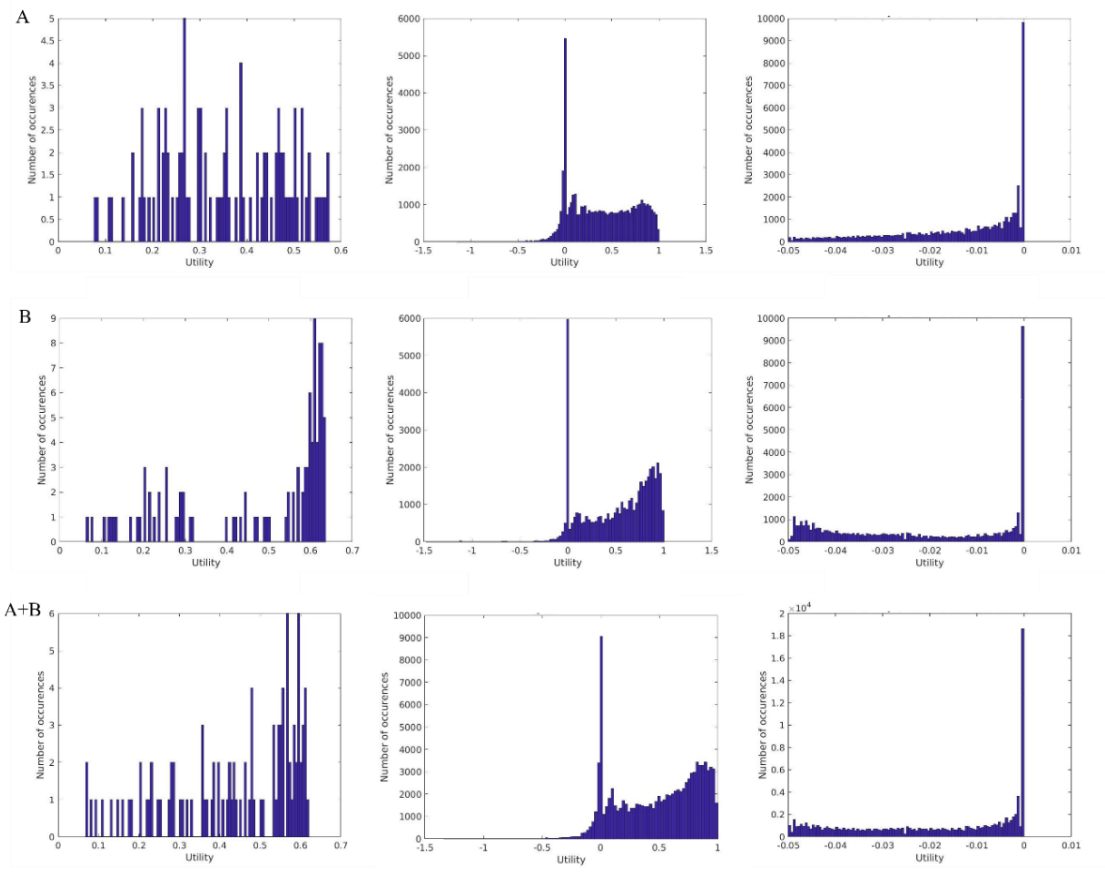

Figura 35. Utilidad obtenida con regresión lineal (Izq: Total; Centro: Pacientes sépticos; Dcha: Pacientes sin sepsis).

En la Figura 35 se muestra la utilidad obtenida para los conjuntos de datos A, B y A+B. Se observa a simple vista que los resultados de utilidad de pacientes sépticos son peores que los mostrados en las Figura 33 y 34, ya que existe un pico muy prominente en cero. En términos de valores máximos, se han dado instancias en que la utilidad ha sido uno, es decir, se ha realizado una predicción perfecta. No obstante, estos casos representan una

minoría frente a la gran concentración de valores alrededor de cero. Esto implica que los valores medios de utilidad para pacientes sépticos son bajos, tal y como se muestra en la . Esta situación, por otro lado, favorece a los pacientes no sépticos, ya que cero es la utilidad óptima en este caso, aunque en términos generales la media obtenida para estos pacientes no es excesivamente superior a la obtenida usando Bayes. Por otro lado, los resultados de utilidad total pueden competir con el clasificador Bayesiano. Los valores máximos obtenidos para los conjuntos de datos A, B y A+B en este caso han sido respectivamente 0,5745, 0,6352 y 0,6195. No obstante, en términos de utilidad media, ésta ha sido notablemente inferior a la obtenida por el clasificador Bayesiano en los tres casos, aunque estos valores medios son comparables a los obtenidos por los primeros clasificados del reto.

Tabla 25. Utilidad media obtenida con regresión lineal.

Conjunto de datos	**Utilidad media**		
	Total	**Pacientes sépticos**	**Pacientes no sépticos**
A	0,352	0,3895	-0,013
B	0,474	0,5164	-0,022
A+B	0,435	0,4828	-0,019

Si se tienen en cuenta las proporciones de utilidad negativa de la regresión lineal mostradas en la , los resultados prueban que este método es claramente inferior a Bayes. Mientras que el clasificador propuesto tenía una proporción P_{E1} inferior al 2,86%, en este caso se supera el 15% en los tres conjuntos de datos. Teniendo en cuenta estas probabilidades junto con los datos de utilidad media calculados, se puede deducir que un clasificador basado en la regresión lineal no es competitivo frente al método Bayesiano.

Tabla 26. Proporción de valores negativos en la función de utilidad usando regresión lineal.

Conjunto de datos	$\boldsymbol{P_{E1}}$
A	0,1978
B	0,1594
A+B	0,1584

5.7.2. Regresión logística

La regresión logística trabaja con probabilidades, al contrario que la regresión lineal, que emplea un rango de valores reales determinados para las variables (apartado 3.7.2).

La regresión logística funciona mejor en aquellos casos en los que se trabaja con variables dicotómicas que no es el caso de las variables clínicas usadas en este trabajo.

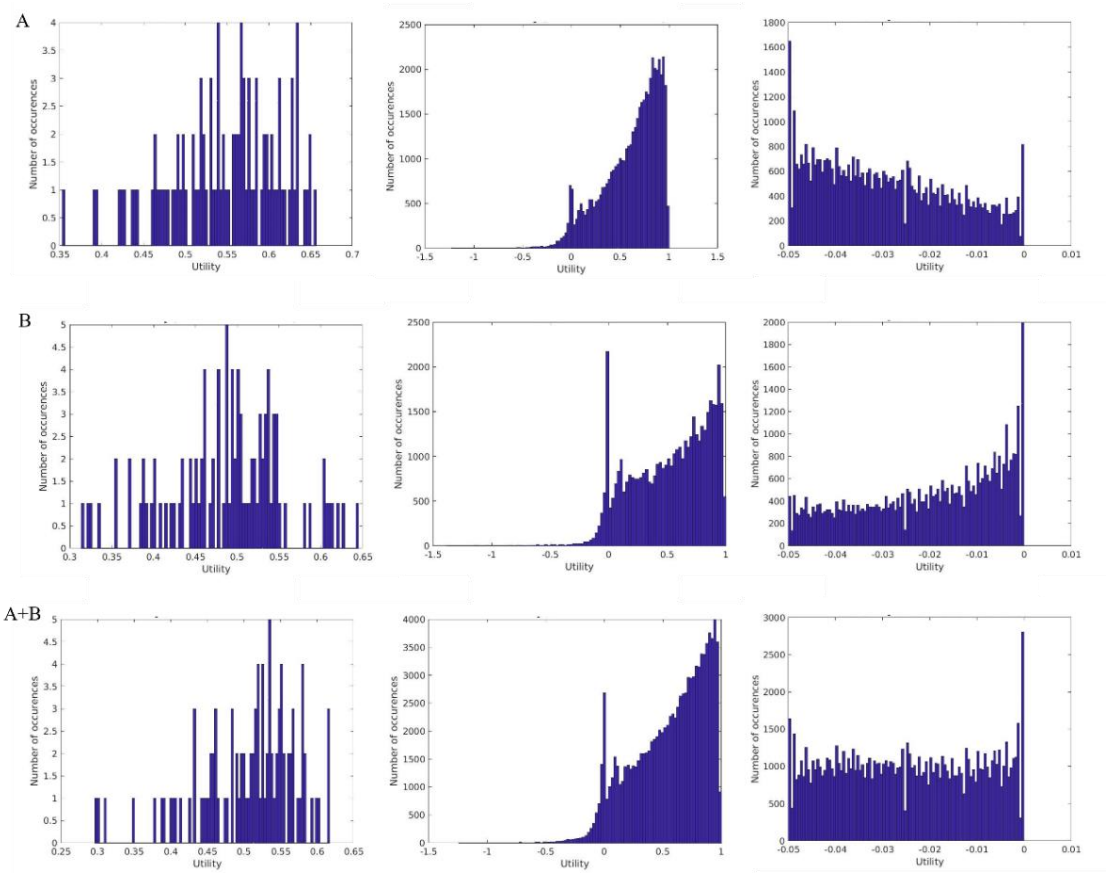

Figura 36. Utilidad obtenida con regresión logística (Izq: Total; Centro: Pacientes sépticos; Dcha: Pacientes sin sepsis).

La distribución de utilidad para pacientes sépticos obtenida, mostrada en la Figura 36 (columna central), es similar a la del clasificador Bayesiano. En los tres casos, dicha utilidad presenta asimetría positiva, especialmente en el caso del conjunto A, en el que el valor de *skewness* es de -0,9025. Los pacientes no sépticos también presentan asimetría positiva, aunque en menor medida, siendo una vez más el conjunto A el que presenta un valor de *skewness* comparable al clasificador Bayesiano (0,4612). El conjunto B no cumple con la asimetría positiva, mientras que en el caso de A+B es casi inexistente, tal y como se muestra en la .

Tabla 27. Valor de skewness para la regresión logística.

Conjunto de datos	Pacientes sépticos	Pacientes no sépticos
A	-0,9025	0,3543

B	-0,5156	-0,3667
A+B	-0,6566	0,0104

En cuanto a la utilidad total, los resultados máximos de cada conjunto de datos son 0,6565 para A, 0,6449 para B y 0,6168 para la combinación de ambos. Una vez más, estos valores son altamente competitivos con los resultados presentados en la . Por otro lado, en la se muestra la utilidad media obtenida en cada caso. Los resultados de la utilidad total y de pacientes sépticos son ligeramente inferiores a los obtenidos por el clasificador Bayesiano, mientras que la utilidad media de los pacientes no sépticos es superior. Esto probablemente se debe a los picos en 0 que se aprecian en las utilidades de ambas clases en la Figura 36. Asimismo, en la se muestran los resultados de la evaluación de calidad del algoritmo. Los resultados son comparables a los obtenidos por el clasificador Bayesiano, siendo la máxima P_{E1} de 8,7% frente al 2,86% obtenido por Bayes. Esto implica que la proporción de valores de utilidad por encima de 0 es ≥ 91,3% (100% - 8,7%), es decir, que la diferencia entre el rendimiento de ambos clasificadores en el peor de los casos es ≤ 5,84%.

Tabla 28. Utilidad media obtenida con regresión logística.

Conjunto de datos	**Utilidad media**		
	Total	**Pacientes sépticos**	**Pacientes no sépticos**
A	0,547	0,6100	-0,029
B	0,486	0,5216	-0,020
A+B	0,505	0,5598	-0,024

Tabla 29. Proporción de valores negativos en la función de utilidad usando regresión logística.

Conjunto de datos	$\boldsymbol{P_{E1}}$
A	0,046
B	0,087
A+B	0,066

5.7.3. SVM

La última técnica que se ha evaluado ha sido SVM. Esta técnica se basa en el cálculo de un hiperplano que permita dividir dos clases de datos de forma óptima en el espacio (apartado 3.7.3).

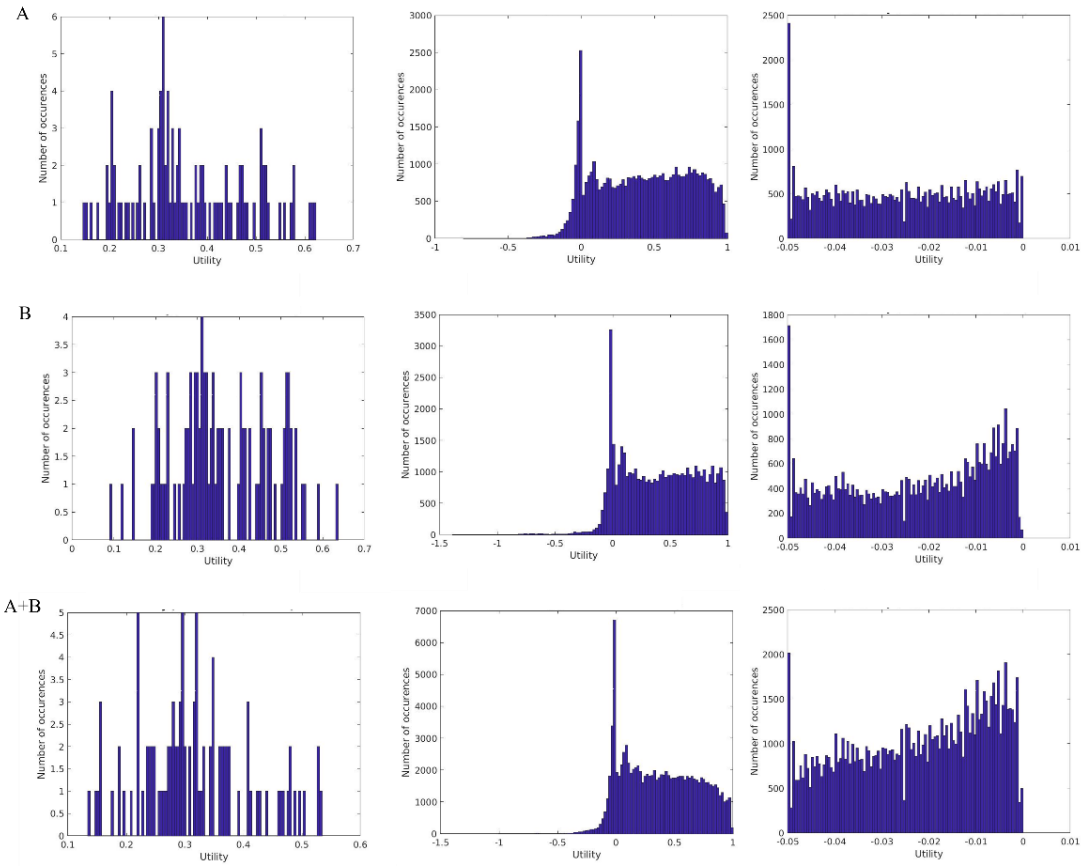

Figura 37. Utilidad obtenida con SVM (Izq: Total; Centro: Pacientes sépticos; Dcha: Pacientes sin sepsis).

Los resultados de utilidad se muestran en la Figura 37. Al igual que ocurría con la regresión lineal, los valores están concentrados alrededor de 0, y el valor máximo de utilidad de pacientes con sepsis es 1, mientras que en el caso de los pacientes sin sepsis es 0. En cuanto a la utilidad total, los resultados máximos obtenidos son similares a los de la regresión lineal: 0,6229 para el conjunto A, 0,6361 para B y 0,5341 para A+B. Si bien estos resultados son ligeramente inferiores al clasificador Bayesiano y la regresión logística, son claramente competitivos con los presentados al reto de PhysioNet.

Al contrario de lo que ocurre con la regresión lineal y logística, el 0 no es el valor de utilidad más frecuente para los pacientes sin sepsis, por lo que la media en este caso es ligeramente peor. De acuerdo con los resultados, que se muestran en la , de entre las tres técnicas analizadas, la utilidad media de SVM es la peor en todos los casos. Además, la

proporción de valores negativos de utilidad de este clasificador () es considerablemente alta en los tres casos, siendo 13,68% el valor más pequeño.

Tabla 30. Utilidad media obtenida con SVM.

Conjunto de datos	**Utilidad media**		
	Total	**Pacientes sépticos**	**Pacientes no sépticos**
A	0,360	0,4209	-0.025
B	0,357	0,4048	-0,023
A+B	0,316	0,3705	-0,022

Tabla 31. Proporción de valores negativos en la función de utilidad usando SVM.

Conjunto de datos	$\boldsymbol{P_{E1}}$
A	0,1368
B	0,1530
A+B	0,1603

En general, tras analizar los resultados obtenidos por los tres clasificadores, se puede concluir que la regresión logística es el único clasificador que puede equipararse al clasificador Bayesiano en los tres aspectos considerados para la comparación (utilidad total máxima, utilidad media y calidad de la función de utilidad). Si bien los otros métodos estudiados en este apartado tienen un rendimiento notable en términos de utilidad total máxima, los valores medios calculados y las probabilidades de error obtenidas indican que, en la mayoría de los casos, no se encuentran a la par con el clasificador de Bayes propuesto.

Capítulo 6. Conclusiones

Este último capítulo está dedicado a las principales conclusiones extraídas de este proyecto. Se han valorado dos aspectos principales: por un lado, si se han logrado los objetivos propuestos para el proyecto; y, por otro lado, la calidad de los resultados finales obtenidos. Finalmente, se indicarán algunas líneas de investigación futuras que surgen de este trabajo.

6.1. Valoración de la consecución de objetivos

El objetivo principal que se ha perseguido en este proyecto ha sido el diseño de un método de monitorización temprana de pacientes con sepsis en Unidades de Cuidados Intensivos a partir de un conjunto de datos clínicos. Para alcanzar dicha meta, se han contemplado dos objetivos específicos:

- O1: Ampliar la funcionalidad del algoritmo de detección temprana para mostrar la evaluación de los pacientes.
- O2: Estudiar técnicas de Machine Learning y, de ser necesario, de aprendizaje profundo que puedan incorporarse al algoritmo.

El primer objetivo se corresponde con el desarrollo de los dos módulos principales del sistema: el módulo de identificación de pacientes sépticos (sección 5.4) y el módulo de detección temprana de eventos de sepsis (sección 5.5). En el primer subsistema se realiza una clasificación de los pacientes dando además unas probabilidades de infección asociadas, comentadas en el apartado 3.6:

- $P\{T|D\}$, es decir, la probabilidad de tener la enfermedad teniendo un test positivo.
- $P\{\overline{T}|\overline{D}\}$, es decir, la probabilidad de no tener la enfermedad teniendo un test negativo.

El uso de estas probabilidades aporta un valor añadido al sistema para los médicos, ya que es un indicador de la evolución de los pacientes en tanto que indica el riesgo de que desarrollen o no la enfermedad a corto plazo. Por otro lado, el segundo módulo se encarga de hacer un seguimiento hora a hora de la sepsis una vez que se ha determinado que un paciente es séptico, indicado si se ha detectado la enfermedad o no. Este sistema es similar al código de alertas implementado en el proyecto BISEPRO [9] [8] en el aspecto de que indica la aparición de sepsis con mayor o menor certeza en cada momento. Por tanto, el primer objetivo se ha alcanzado.

En cuanto al segundo objetivo, se han evaluado tres técnicas de Machine Learning alternativas al clasificador bayesiano: regresión lineal, regresión logística y SVM. De acuerdo con los resultados obtenidos, la regresión logística tiene un rendimiento equiparable al método propuesto en este trabajo, y podría ser potencialmente incorporada al algoritmo con fin de mejorar los resultados de Bayes, o bien ser usada como sistema alternativo. La regresión lineal y SVM, por su parte, no han alcanzado el nivel de utilidad del clasificador Bayesiano.

No se han analizado técnicas de aprendizaje profundo, ya que, de acuerdo con varios autores [49][50], lo recomendable es comenzar por modelos de inteligencia artificial convencionales y utilizar el aprendizaje profundo únicamente cuando estas técnicas no sean suficiente para resolver el problema analizado. Para el caso presentado en este trabajo, se ha demostrado que tanto el clasificador bayesiano como la regresión logística son "clasificadores base" capaces de dar resultados científicamente competitivos para el problema de la detección temprana de la sepsis. Por otro lado, utilizar estrategia de combinación de clasificadores (base) y un posprocesado basado en probabilidades acumuladas aportan mejoras muy significativas. En consecuencia, no parece que incorporar métodos de aprendizaje profundo llegue a aportar una mejora sustancial del sistema propuesto. Por otro lado, la forma en que se hace la combinación de clasificadores permite dar una solución muy buena al problema de datos perdidos o aberrantes.

Como conclusión, a través de la finalización de ambos objetivos específicos se ha logrado conseguir el objetivo principal del este trabajo.

6.2. Valoración de los resultados

Este trabajo toma como referencia el reto "Early Prediction of Sepsis from Clinical Data - the PhysioNet Computing in Cardiology Challenge 2019" [14]. Al compartir unos objetivos y una base de datos comunes, este reto representa un marco ideal para la comparación de metodologías en la comunidad científica. Como se ha expuesto en el capítulo anterior, los resultados obtenidos por el sistema propuesto en este proyecto superan con creces a los ocho primeros clasificados del reto de 2019. Mientras que los mejores resultados del reto no alcanzaron el 0,4 de utilidad, el algoritmo presentado en este trabajo ha conseguido resultados mayores que 0,6, superando también lo conseguido en el trabajo previo realizado por Alba Manso, a partir del cual se ha desarrollado el sistema actual.

Modelo empleado	Utilidad
Sistema de monitorización propuesto – Conjunto A	0,675
Sistema de monitorización propuesto – Conjunto A+B	0,641
Sistema de monitorización propuesto – Conjunto B	0,621
Algoritmo para la identificación temprana de la sepsis	0,522
Modelo de regresión basado en la firma de característica	0,360
Redes neuronales	0,345
Conjunto de modelos XGBoost	0,339
Modelo Time-Phased	0,337
Modelo basado en XGBoost	0,337
Árboles de decisión basados en refuerzo	0,332
Modelo basado en XGBoost	0,331
Arquitectura neuronal	0,328

Hay que destacar que la puntuación final del reto fue calculada sobre la población de datos total conformada por los tres conjuntos de datos, A, B y C [14]. Por tanto, los experimentos no se han podido llevar a cabo en un entorno totalmente fiel al del reto original debido a que los autores del mismo no proporcionan acceso al conjunto de datos C. El método empleado para equilibrar las condiciones de experimentación ha sido mezclar los dos conjuntos de datos públicos. De esta forma, la utilidad obtenida para esta combinación A+B ha sido la segunda más alta de los experimentos realizados (0,641). Este hecho, junto con la selección aleatoria de pacientes, ha permitido validar la metodología usada en multitud de escenarios.

Otro aspecto a tener en cuenta es que, aunque las métricas de la combinación de las estrategias producto y *majority voting* no parezcan muy buenas *a priori*, la aplicación de este método al cálculo de la utilidad ha demostrado lo contrario. Concretamente, se ha obtenido una utilidad máxima de 0,617 para el conjunto de datos A, 0,6028 para el conjunto de datos B y 0,593 para la mezcla de ambos conjuntos. Por otro lado, la utilidad media ha sido de 0,5709 para el conjunto de datos A, 0,5201 para el conjunto de datos B y 0,5309 para el conjunto A+B. Dichos resultados se pueden consultar en el Anexo III. Estos resultados son ligeramente inferiores a los conseguidos usando únicamente *majority voting*. No obstante, son valores muy superiores a los conseguidos por los primeros

clasificados del reto y comparables al mejor resultado obtenido por Alba Manso en su trabajo. En definitiva, los resultados prueban que el sistema desarrollado en este TFM es muy competitivo.

6.3. Líneas futuras

Si bien la base de datos de Physionet es muy buena, de cara al futuro se puede plantear aplicar el diseño propuesto en este proyecto a una base de datos propia de hospitales de Canarias, la Comunidad Valenciana y otros sitios de España. El objetivo fundamental de esto es avanzar hacia un sistema capaz de trabajar en tiempo real con pacientes en UCI. De la mano de esta línea de trabajo está conseguir datos más ricos (comorbilidades, etc.), periodos de muestreo menores a 60 minutos (p.ej., datos cada 10) y, a poder ser, menor cantidad de datos perdidos o datos aberrantes.

Dado que el problema de datos perdidos podría ser inevitable, otra línea futura es hacer una labor de investigación con métodos de imputación de datos que rellenen vacíos con la mayor fidelidad posible a los datos originales.

Un problema que subyace en todo el proyecto es el de predicción a futuro, esto es, ser capaces de predecir cuál será el estado de sepsis del paciente dentro de 1, 2, …, *n* horas, o incluso días. A este respecto, quedan muchas técnicas de Machine Learning convencional que explorar, como los Modelos Ocultos de Markov. No obstante, dado el potencial que presentan las técnicas de Deep Learning para hacer predicciones con series temporales, no debe descartarse su uso, pues puede abrir un nuevo abanico de posibilidades.

Por último, queda el paso de la investigación a la transferencia. En este sentido se potenciará la colaboración con hospitales de Canarias y la Comunidad Valenciana.

Referencias

[1] M. Singer *et al.*, "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, Feb. 2016, doi: 10.1001/jama.2016.0287. [En línea]. Disponible: https://doi.org/10.1001/jama.2016.0287

[2] G. Polat, R. A. Ugan, E. Cadirci, and Z. Halici, "Sepsis and Septic Shock: Current Treatment Strategies and New Approaches," *Eurasian J. Med.*, vol. 49, no. 1, pp. 53–58, Feb. 2017, doi: 10.5152/eurasianjmed.2017.17062. [En línea]. Disponible: https://pubmed.ncbi.nlm.nih.gov/28416934

[3] "Sepsis — Global Sepsis Alliance." [En línea]. Disponible: https://www.global-sepsis-alliance.org/sepsis. [Accedido: 14-Sep-2021]

[4] "Sepsis." [En línea]. Disponible: https://www.who.int/news-room/fact-sheets/detail/sepsis. [Accedido: 14-Sep-2021]

[5] C. J. Paoli, M. A. Reynolds, M. Sinha, M. Gitlin, and E. Crouser, "Epidemiology and Costs of Sepsis in the United States-An Analysis Based on Timing of Diagnosis and Severity Level," *Crit. Care Med.*, vol. 46, no. 12, pp. 1889–1897, Dec. 2018, doi: 10.1097/CCM.0000000000003342. [En línea]. Disponible: https://pubmed.ncbi.nlm.nih.gov/30048332

[6] "En España se registran unos 50.000 casos de sepsis al año, aunque podría ser una cifra 4 o 5 veces inferior a la real," 2021. [En línea]. Disponible: https://www.infosalus.com/salud-investigacion/noticia-espana-registran-50000-casos-sepsis-ano-podria-ser-cifra-veces-inferior-real-20210913135544.html. [Accedido: 14-Sep-2021]

[7] "El Código Sepsis reduce la mortalidad por debajo del 20%." [En línea]. Disponible: https://www.consalud.es/pacientes/dias-mundiales/codigo-sepsis-reduce-mortalidad-20-hospitales_102090_102.html. [Accedido: 14-Sep-2021]

[8] "El IIC colabora en el Proyecto BISEPRO para detectar sepsis - IIC." [En línea]. Disponible: https://www.iic.uam.es/noticias/el-iic-colabora-en-el-proyecto-bisepro-para-detectar-sepsis/. [Accedido: 29-Sep-2021]

[9] E. Santa María, “BISEPRO: inteligencia artificial para agilizar el diagnóstico de la sepsis - iSanidad,” 07-Nov-2018. [En línea]. Disponible: https://isanidad.com/126725/bisepro-inteligencia-artificial-para-agilizar-el-diagnostico-de-la-sepsis/. [Accedido: 30-Sep-2021]

[10] S. M. Lauritsen *et al.*, “Early detection of sepsis utilizing deep learning on electronic health record event sequences,” *Artif. Intell. Med.*, vol. 104, no. January, p. 101820, 2020, doi: 10.1016/j.artmed.2020.101820. [En línea]. Disponible: https://doi.org/10.1016/j.artmed.2020.101820

[11] I. Stanculescu, C. K. I. Williams, and Y. Freer, “Autoregressive Hidden Markov Models for the Early Detection of Neonatal Sepsis,” *IEEE J. Biomed. Heal. Informatics*, vol. 18, no. 5, pp. 1560–1570, 2014, doi: 10.1109/JBHI.2013.2294692.

[12] A. Manso Fernández, “Algoritmo para la identificación temprana de la sepsis a partir de datos clínicos,” Universidad de Alicante, 2020.

[13] “PhysioNet.” [En línea]. Disponible: https://physionet.org/. [Accedido: 15-Sep-2021]

[14] M. Reyna *et al.*, “Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019,” *2019 Comput. Cardiol. Conf.*, vol. 45, pp. 2019–2022, 2019, doi: 10.22489/cinc.2019.412.

[15] J. S. H. Botero and M. C. F. Pérez, “The history of sepsis from ancient Egypt to the XIX century,” in *Sepsis-an ongoing and significant challenge*, Intechopen, 2012.

[16] D. J. J. Muckart and S. Bhagwanjee, “American College of Chest Physicians/Society of Critical Care Medicine Consensus Conference definitions of the systemic inflammatory response syndrome and allied disorders in relation to critically injured patients,” *Crit. Care Med.*, vol. 25, no. 11, pp. 1789–1795, 1997.

[17] M. M. Levy *et al.*, “2001 sccm/esicm/accp/ats/sis international sepsis definitions conference,” *Intensive Care Med.*, vol. 29, no. 4, pp. 530–538, 2003.

[18] R. R. Valero, “Valoración de la gravedad, estratificación y predicción en el enfermo con sepsis grave,” *Rev. Electrónica Med. Intensiva*, vol. 5, no. 3, 2005.

[19] M. Shapiro, “Escalas pronósticas en la Unidad de Terapia Intensiva,” *Rev. la Asoc. Mex. Med. Crítica y Ter. Intensiva*, vol. 26, no. 4, pp. 234–241, 2012 [En línea]. Disponible: www.medigraphic.org.mx

[20] R. A. Balk, “Systemic inflammatory response syndrome (SIRS): Where did it come from and is it still relevant today?,” *Virulence*, vol. 5, no. 1, pp. 20–26, 2014, doi: 10.4161/viru.27135.

[21] A. Khojandi, V. Tansakul, X. Li, R. S. Koszalinski, and W. Paiva, “Prediction of Sepsis and In-Hospital Mortality Using Electronic Health Records,” *Methods Inf. Med.*, vol. 57, no. 4, pp. 185–193, 2018, doi: 10.3414/ME18-01-0014.

[22] F. Lombi and H. Trimarchi, “Nuevas definiciones de Injuria Renal Aguda y sepsis: impacto en el abordaje diagnóstico,” 20-Dec-2016. [En línea]. Disponible: https://www.revistarenal.org.ar/index.php/rndt/article/download/144/493?inline=1. [Accedido: 03-Apr-2022]

[23] S. Haydar, M. Spanier, P. Weems, S. Wood, and T. Strout, “Comparison of QSOFA score and SIRS criteria as screening mechanisms for emergency department sepsis,” *Am. J. Emerg. Med.*, vol. 35, no. 11, pp. 1730–1733, 2017, doi: 10.1016/j.ajem.2017.07.001. [En línea]. Disponible: https://doi.org/10.1016/j.ajem.2017.07.001

[24] E. P. Raith *et al.*, “Prognostic Accuracy of the SOFA Score, SIRS Criteria, and qSOFA Score for In-Hospital Mortality Among Adults With Suspected Infection Admitted to the Intensive Care Unit,” *JAMA*, vol. 317, no. 3, pp. 290–300, Jan. 2017, doi: 10.1001/JAMA.2016.20328. [En línea]. Disponible: https://jamanetwork.com/journals/jama/fullarticle/2598267. [Accedido: 11-Jan-2022]

[25] D. T. Wong and W. A. Knaus, “Predicting outcome in critical care: the current status of the APACHE prognostic scoring system,” *Can. J. Anaesth.*, vol. 38, no. 3, pp. 374–383, 1991, doi: 10.1007/BF03007629.

[26] G. Firman, “Sistema de Clasificación de Severidad de Enfermedad APACHE II – MedicalCRITERIA.com,” 20-Feb-2009. [En línea]. Disponible: https://medicalcriteria.com/web/es/utiapache/. [Accedido: 03-Apr-2022]

[27] S. Gien J, “Valor predictivo de la escala APACHE II sobre la mortalidad en una

unidad de cuidados intensivos de adultos en la ciudad de Mérida Yucatán," *Med. Crit. y Ter. intensiva*, vol. 20, p. 1, 2006.

[28] C. Chatzicostas *et al.*, "Comparison of Ranson, APACHE II and APACHE III scoring systems in acute pancreatitis," *Pancreas*, vol. 25, no. 4, pp. 331–335, 2002.

[29] F. Sadaka, C. EthmaneAbouElMaali, M. A. Cytron, K. Fowler, V. M. Javaux, and J. O'Brien, "Predicting Mortality of Patients With Sepsis: A Comparison of APACHE II and APACHE III Scoring Systems," *J. Clin. Med. Res.*, vol. 9, no. 11, pp. 907–910, Nov. 2017, doi: 10.14740/jocmr3083w. [En línea]. Disponible: https://pubmed.ncbi.nlm.nih.gov/29038667

[30] A. G. Rapsang and D. C. Shyam, "Scoring systems in the intensive care unit: a compendium," *Indian J. Crit. care Med. peer-reviewed, Off. Publ. Indian Soc. Crit. Care Med.*, vol. 18, no. 4, p. 220, 2014.

[31] R. Gauer, "Early recognition and management of sepsis in adults: the first six hours," *Am. Fam. Physician*, vol. 88, no. 1, pp. 44–53, 2013.

[32] E. Rivers *et al.*, "Early goal-directed therapy in the treatment of severe sepsis and septic shock," *N. Engl. J. Med.*, vol. 345, no. 19, pp. 1368–1377, 2001.

[33] M. Moor, B. Rieck, M. Horn, and C. R. Jutzeler, "Early Prediction of Sepsis in the ICU Using Machine Learning : A Systematic Review," vol. 8, no. May, 2021, doi: 10.3389/fmed.2021.607952.

[34] H. J. Kam and H. Y. Kim, "Learning representations for the early detection of sepsis with deep neural networks," *Comput. Biol. Med.*, vol. 89, no. August, pp. 248–255, 2017, doi: 10.1016/j.compbiomed.2017.08.015. [En línea]. Disponible: https://doi.org/10.1016/j.compbiomed.2017.08.015

[35] J. S. Calvert *et al.*, "A computational approach to early sepsis detection," *Comput. Biol. Med.*, vol. 74, pp. 69–73, 2016, doi: 10.1016/j.compbiomed.2016.05.003. [En línea]. Disponible: http://dx.doi.org/10.1016/j.compbiomed.2016.05.003

[36] J. Morrill *et al.*, "The Signature-based Model for Early Detection of Sepsis from Electronic Health Records in the Intensive Care Unit," pp. 2–5.

[37] M. Zabihi, S. Kiranyaz, and M. Gabbouj, "Sepsis Prediction in Intensive Care Unit Using Ensemble of XGboost Models," *2019 Comput. Cardiol. Conf.*, vol. 45, pp.

27–30, 2019, doi: 10.22489/cinc.2019.238.

[38] X. Li, Y. Kang, X. Jia, J. Wang, and G. Xie, “TASP : A Time-Phased Model for Sepsis Prediction,” vol. 46, pp. 1–4, 2019, doi: 10.22489/CinC.2019.049.

[39] J. Singh, K. Oshiro, R. Krishnan, M. Sato, T. Ohkuma, and N. Kato, “Utilizing Informative Missingness for Early Prediction of Sepsis,” vol. 46, pp. 1–4, 2019, doi: 10.22489/CinC.2019.280.

[40] I. Hammoud, IV Ramakrishnan1, and M. Henry, “Early Prediction of Sepsis using Gradient Boosting Decision Trees With Custom Label Weighting,” pp. 2–4, 2019.

[41] Y. Chang, J. Rubin, G. Boverman, S. Vij, A. Rahman, and A. Natarajan, “A Multi-Task Imputation and Classification Neural Architecture for Early Prediction of Sepsis from Multivariate Clinical Time Series,” vol. 46, no. 1, pp. 2–5, 2019, doi: 10.22489/CinC.2019.110.

[42] S. Brown, “Machine learning, explained | MIT Sloan,” 21-Apr-2021. [En línea]. Disponible: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained. [Accedido: 24-Mar-2022]

[43] T. Jiang, J. L. Gradus, and A. J. Rosellini, “Supervised Machine Learning: A Brief Primer,” *Behavior Therapy*, vol. 51, no. 5. pp. 675–687, 2020.

[44] B. Ammanath, S. Hupfer, and D. Jarvis, “Thriving in the era of pervasive AI,” *Deloitte Insights*, pp. 1–25, 2020.

[45] P. P. Shinde and S. Shah, “A Review of Machine Learning and Deep Learning Applications,” *Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018*, 2018, doi: 10.1109/ICCUBEA.2018.8697857.

[46] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[47] M. Helmstaedter, K. L. Briggman, S. C. Turaga, V. Jain, H. S. Seung, and W. Denk, “Connectomic reconstruction of the inner plexiform layer in the mouse retina,” *Nature*, vol. 500, no. 7461, pp. 168–174, 2013.

[48] M. K. K. Leung, H. Y. Xiong, L. J. Lee, and B. J. Frey, “Deep learning of the tissue-regulated splicing code,” *Bioinformatics*, vol. 30, no. 12, pp. 121–129, 2014,

doi: 10.1093/bioinformatics/btu277.

[49] V. Ulhmann, L. Donati, and D. Sage, “A Practical Guide to Supervised Deep Learning for Bioimage Analysis: Challenges and good practices,” *IEEE Signal Process. Mag.*, vol. 39, no. March, pp. 73–86, 2022.

[50] C. Jutten, “Smart Deep Learning for Biological Images and Signals,” *IEEE Signal Process. Mag.*, vol. 39, no. 2, p. 13, 2022.

[51] M. Batta, “Machine Learning Algorithms - A Review ,” *Int. J. Sci. Res. (IJ*, vol. 9, no. 1, pp. 381-undefined, 2020, doi: 10.21275/ART20203995.

[52] J. Tello Cáceres, “Reconocimiento de patrones y el aprendizaje no supervisado,” *Esc. Técnica Super. Informática Univ. Alcalá*, no. June, p. 738, 2006 [En línea]. Disponible: http://www-etsi2.ugr.es/depar/ccia/rf/www/tema1_00-01_www/node6.html.

[53] J. A. Hernández, “Métodos de reducción de dimensionalidad: Análisis comparativo de los métodos APC, ACPP y ACPK,” *Uniciencia*, vol. 30, no. 1, pp. 115–122, 2016.

[54] Z. Ghahramani, “Unsupervised learning,” in *Summer School on Machine Learning*, 2003, pp. 72–112.

[55] L. P. Kaelbling, M. L. Littman, and A. W. Moore, “Reinforcement Learning : A Survey,” pp. 237–285, 1996.

[56] B. A. Olshausen, “Bayesian probability theory,” *Redw. Cent. Theor. Neurosci. Helen Wills Neurosci. Inst. Univ. Calif. Berkeley, Berkeley, CA*, 2004.

[57] O. D. Castrillón, J. A. Giraldo, and W. Sarache, “Sistema de clasificación Bayesiano basado en múltiples clases,” *Sist. Cibernética e Informática*, 2008.

[58] S. D. Benning, “Clinical Statistics , Psychodiagnosis , and Bayes ’ Theorem,” vol. 2008, no. December, 2016.

[59] S. Taheri and M. Mammadov, “Learning the naive Bayes classifier with optimization models,” *Int. J. Appl. Math. Comput. Sci.*, vol. 23, no. 4, 2013.

[60] S. Ray, “Commonly Used Machine Learning Algorithms | Data Science,” 2017. [En línea]. Disponible: https://www.analyticsvidhya.com/blog/2017/09/common-

machine-learning-algorithms/. [Accedido: 22-Jan-2022]

[61] N. J. Gogtay and U. M. Thatte, “Statistical evaluation of diagnostic tests (part 1): Sensitivity, specificity, positive and negative predictive values,” *J. Assoc. Physicians India*, vol. 65, no. JUNE, pp. 80–84, 2017.

[62] H. Kang, “Statistical Evaluation for Medical Screening & Diagnostic Tests,” 2020.

[63] X. Ying, “An Overview of Overfitting and its Solutions,” *J. Phys. Conf. Ser.*, vol. 1168, no. 2, 2019, doi: 10.1088/1742-6596/1168/2/022022.

[64] D. Berrar, “Cross-validation,” vol. 1, pp. 542–545, 2018.

[65] L. I. Kuncheva, *Combining Pattern Classifiers*, Second edi. Wiley, 2004 [En línea]. Disponible: https://En línealibrary.wiley.com/doi/book/10.1002/0471660264

[66] M. P. Fewell and J. M. Thredgold, “Cumulative Track-Initiation Probability as a Basis for Assessing Sonar-System Performance in Anti-Submarine Warfare,” 2009.

[67] A. K. Teng and A. B. Wilcox, “A review of predictive analytics solutions for sepsis patients,” *Appl. Clin. Inform.*, vol. 11, no. 03, pp. 387–398, 2020.

[68] J. Dagnino, “Regresión lineal,” *Rev. Chil. Anest.*, vol. 43, no. 2, 2014.

[69] M. C. Carollo Limeres, “Regresión lineal simple,” pp. 1–31, 2016 [En línea]. Disponible: http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_50140116_Regr_ simple_2011_12.pdf

[70] I. M. Peláez, “Modelos de regresión: lineal simple y regresión logística,” *Rev. Seden*, vol. 14, pp. 195–214, 2016.

[71] F. Jaimes, J. Farbiarz, D. Alvarez, and C. Martínez, “Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room,” *Crit. Care*, vol. 9, no. 2, p. R150, 2005, doi: 10.1186/cc3054. [En línea]. Disponible: https://doi.org/10.1186/cc3054

[72] “nRF52840 Product Specification.” [En línea]. Disponible: https://infocenter.nordicsemi.com/pdf/nRF52840_PS_v1.5.pdf. [Accedido: 24-Oct-2021]

[73] Rushikesh Pupale, “Support Vector Machines(SVM) — An Overview,” 2018. [En línea]. Disponible: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989. [Accedido: 28-Mar-2022]

[74] D. L. Whaley III, “The interquartile range: Theory and estimation,” East Tennessee State University, Ann Arbor, 2005 [En línea]. Disponible: https://www.proquest.com/dissertations-theses/interquartile-range-theory-estimation/docview/304994791/se-2?accountid=14705

[75] “Support Vector Machine (SVM) - MATLAB & Simulink.” [En línea]. Disponible: https://es.mathworks.com/discovery/support-vector-machine.html. [Accedido: 28-Mar-2022]

[76] J. Guillén *et al.*, “Predictive models for severe sepsis in adult ICU patients,” *2015 Syst. Inf. Eng. Des. Symp. SIEDS 2015*, no. c, pp. 182–187, 2015, doi: 10.1109/SIEDS.2015.7116970.

[77] E. Originale, “Understanding receiver operating characteristic (ROC) curves,” vol. 8, no. 1, pp. 19–20, 2006.

[78] A. Díez Herranz and M. Tobal González, “Las curvas ROC en la evaluación de las pruebas diagnósticas.,” *Med. Clin. (Barc).*, vol. 108, no. 1, pp. 34–35, 1997.

[79] N. J. Gogtay and U. M. Thatte, “Statistical evaluation of diagnostic tests – part 2 [Pre-test and post-test probability and odds, likelihood ratios, receiver operating characteristic curve, youden’s index and diagnostic test biases],” *J. Assoc. Physicians India*, vol. 65, no. JULY, pp. 86–91, 2017.

[80] I. Markoulidakis, I. Rallis, I. Georgoulas, G. Kopsiaftis, A. Doulamis, and N. Doulamis, “Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem,” *Technologies*, vol. 9, no. 4, p. 81, 2021.

[81] “Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019 v1.0.0.” [En línea]. Disponible: https://physionet.org/content/challenge-2019/1.0.0/. [Accedido: 07-Apr-2022]

[82] D. Ghosh and A. Vogt, “Outliers: An Evaluation of Methodologies,” *Jt. Stat. Metings*, pp. 3455–3460, 2012.

[83] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, “A targeted real-time early warning score (TREWScore) for septic shock,” *Sci. Transl. Med.*, vol. 7, no. 299, 2015, doi: 10.1126/scitranslmed.aab3719.

[84] C. Begoña, “Contraste de hipótesis nula y alternativa,” pp. 1–14, 2017 [En línea]. Disponible: http://diposit.ub.edu/dspace/bitstream/2445/117643/1/Phipotesis-apuntes1718-DD.pdf

[85] A. R. Julio, “Wilcoxon-Mann-Whitney test como alternativa al t-test,” pp. 1–18, 2017.

[86] S. P. Lloyd, “Least Squares Quantization in PCM,” vol. I, no. 2, pp. 129–137, 1982.

[87] M. Beltrán Pascual, A. Muñoz Martínez, and Á. Muñoz Alamillos, “Bayesian networks applied to credit scoring problems. A practical application,” *Cuad. Econ.*, vol. 37, no. 104, pp. 73–86, 2014, doi: 10.1016/j.cesjef.2013.07.001.

[88] “Asimetría y curtosis.” [En línea]. Disponible: https://www.universoformulas.com/estadistica/descriptiva/asimetria-curtosis/. [Accedido: 16-May-2022]

Parte II: Presupuesto

Presupuesto detallado

Debido a que el Colegio Oficial de Ingenieros de Telecomunicación (COIT) no cuenta con recomendaciones propias para el cómputo de un presupuesto, este apartado recoge el presupuesto del proyecto propuesto conforme a las recomendaciones del Colegio Oficial de Ingenieros Técnicos de Telecomunicación (COITT). En dichas recomendaciones se contemplan los siguientes apartados:

1. Recursos materiales.
2. Trabajo tarifado por tiempo empleado.
3. Costes de redacción del trabajo de fin de máster.
4. Derechos de visado.
5. Gastos de tramitación y envío.
6. Aplicación de impuestos.

P1. Recursos materiales

Dentro de los recursos materiales se incluyen los recursos *software* y *hardware* utilizados para la elaboración del proyecto. A su vez, los recursos *software* comprenden el paquete de ofimática usado para la escritura de la memoria y el *software* de desarrollo del algoritmo de monitorización, mientras que los recursos *hardware* consisten en el ordenador en el que se ha llevado a cabo el desarrollo del sistema.

Se estipula el coste de amortización para un período de cuatro años. Se ha utilizado un sistema de amortización lineal, es decir, se supone que el material inmovilizado se desvaloriza de forma constante a lo largo del período indicado. El coste de amortización viene dado por la siguiente expresión:

$$Coste\ de\ amortización = \frac{Valor\ de\ adquisición - Valor\ residual}{Años\ de\ vida\ útil}$$

El trabajo de fin de máster ha tenido una duración de cuatro meses en total, que es un período inferior a los cuatro años de amortización estipulados. Por tanto, los costes que se presentan en esta memoria son los derivados de los 4 meses durante los que se ha desarrollado este proyecto.

P1.1. Recursos *software*

Las herramientas *software* que se han usado son:

- *Software* de procesado de datos MATLAB.
- Paquete de ofimática Microsoft Office 365.

La Universidad de Las Palmas de Gran Canaria provee licencias académicas de ambas herramientas a los estudiantes, por lo que no ha sido necesario realizar una compra de licencias.

Por tanto, el coste total de los recursos *software* es de **cero euros (0€).**

P1.2. Recursos *hardware*

Los recursos *hardware* usados en este proyecto consisten únicamente en un ordenador de desarrollo:

- Ordenador portátil Acer modelo N20C1.

Por lo que el coste de los recursos *hardware* es el siguiente:

Descripción	Valor de adquisición	Valor residual	Coste de amortización
Ordenador portátil	800€	200€	150€

El coste total de los recursos *hardware* asciende a **ciento cincuenta euros (150,00€).**

P2. Trabajo tarifado por tiempo empleado

En este trabajo se han empleado un total de 300 horas para la búsqueda de información, estudio de soluciones, desarrollo e implementación del sistema y redacción de documentación. Siguiendo las recomendaciones establecidas por el COIT, el importe de las horas trabajadas se calcula con la siguiente fórmula:

$$H = C_t \cdot (74{,}88 \cdot H_n) + C_t \cdot (96{,}72 \cdot H_e)$$

Donde:

- H indica los honorarios totales por tiempo dedicado.
- H_n indica los honorarios en jornada laboral normal.
- H_e indica los honorarios fuera de la jornada laboral normal.
- C_t indica el factor de corrección según el número de horas trabajadas.

Según lo establecido, el factor de corrección C_t varía dependiendo del número de horas empleadas. En este caso, para 300 horas trabajadas el factor de corrección es de 0,60.

Teniendo en cuenta que no se ha trabajado fuera del horario laboral normal, los honorarios totales por tiempo dedicado al trabajo ascienden a **trecemil cuatrocientos setenta y ocho euros y cuarenta céntimos (13.478,40€).**

P3. Costes de redacción del trabajo de fin de máster

El coste de la redacción del proyecto viene dado por la siguiente expresión:

$$R = 0{,}07 \cdot P \cdot C_n$$

Donde P es el presupuesto del proyecto y C_n es un coeficiente de ponderación que depende del presupuesto.

El presupuesto total calculado hasta este momento asciende a 13.628,40€. Según lo recomendado, para presupuesto inferior a 30.050,00€, el coeficiente de ponderación tiene valor 1,00. Por tanto, el coste derivado de la redacción del proyecto es de **novecientos cincuenta y tres euros con noventa y nueve céntimos (953,99€)**.

P4. Derechos de visado

Los gastos de visado del proyecto según el COIT se calculan con la siguiente expresión:

$$V = 0{,}0035 \cdot P \cdot C_v$$

Donde P es el presupuesto del proyecto y C_v es un coeficiente de reducción que depende del presupuesto.

El valor del presupuesto es de 14.582,39€, por lo que según la recomendación establecida el coeficiente de reducción tiene valor 1,00 al no superar la cifra de 30.050,00€. Así, el coste de los derechos de visado del proyecto es de **cincuenta y un euros con cuatro céntimos (51,04€).**

P5. Gastos de tramitación y envío

Los gastos de tramitación y envío tiene un precio fijo de **seis euros y un céntimo (6,01€).**

P6. Aplicación de impuestos

El coste total del proyecto antes de aplicar impuestos asciende a 14.639,18€. Tras aplicar el Impuesto General Indirecto Canario (IGIC) el coste final es el siguiente:

Recursos	Coste
Recursos materiales	150,00€
Trabajo tarifado por tiempo empleado	13.478,40€
Redacción del TFM	953,99€
Derechos de visado del COIT	51,04€
Gastos de tramitación y envío	6,01€
Subtotal	14.639,18€
Aplicación de impuestos (IGIC 7%)	1.024,74€
Total	**15.663,92€**

Por tanto, el presupuesto total de este trabajo de fin de máster es de **quince mil seiscientos sesenta y tres euros con noventa y dos céntimos (15.663,92€).**

Las Palmas de Gran Canaria a 14 de julio de 2022.

Firma:

Parte III: Anexos

Anexo I: Matriz de confusión para la identificación de eventos de sepsis

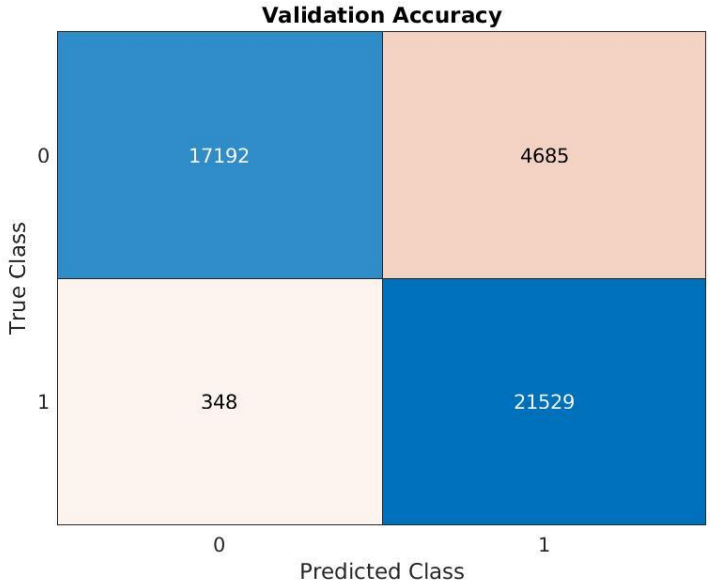

Figura 38. Matriz de confusión de identificación de eventos de sepsis para el conjunto de datos A.

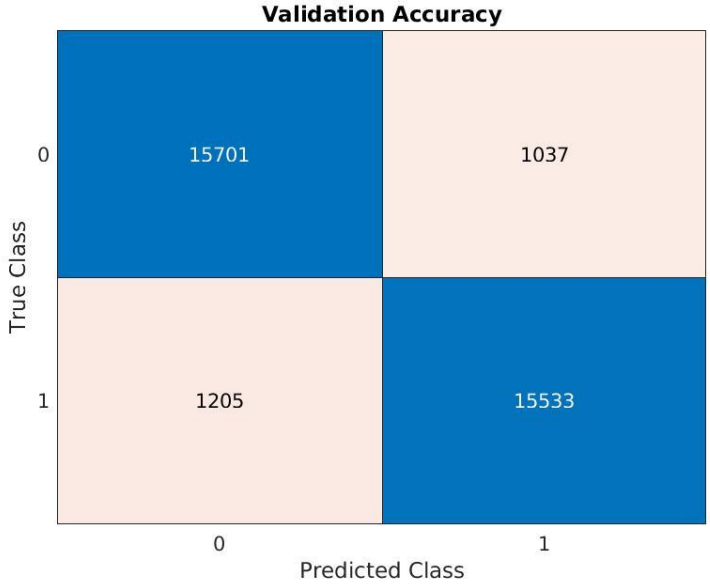

Figura 39. Matriz de confusión de identificación de eventos de sepsis para el conjunto de datos B.

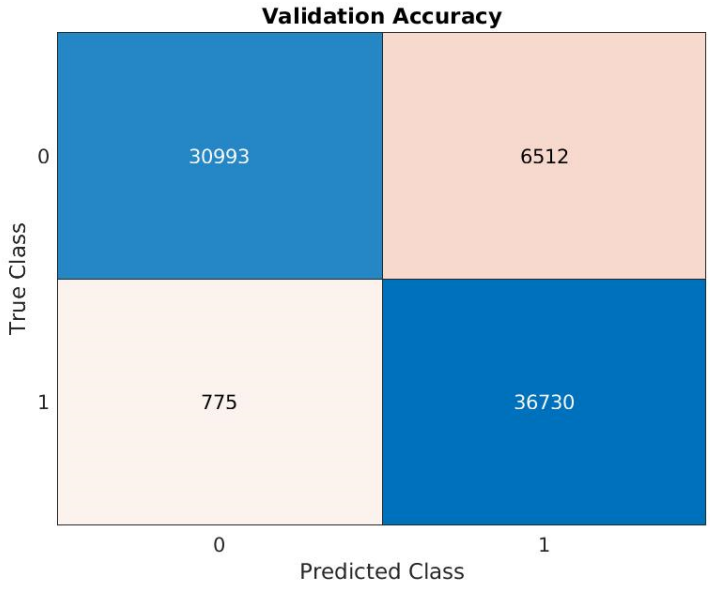

Figura 40. Matriz de confusión de identificación de eventos de sepsis para el conjunto de datos A+B.

Anexo II: Utilidad para un mínimo de 48 horas

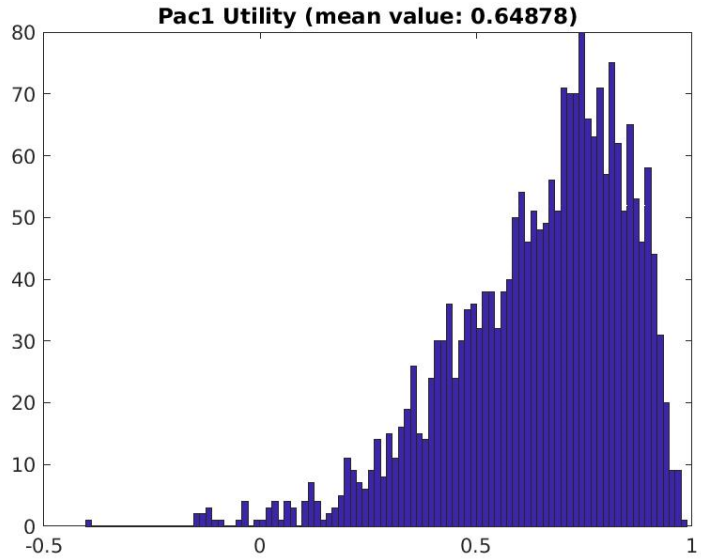

Figura 41. Utilidad para pacientes sépticos para un mínimo de 48 horas en el conjunto de datos A.

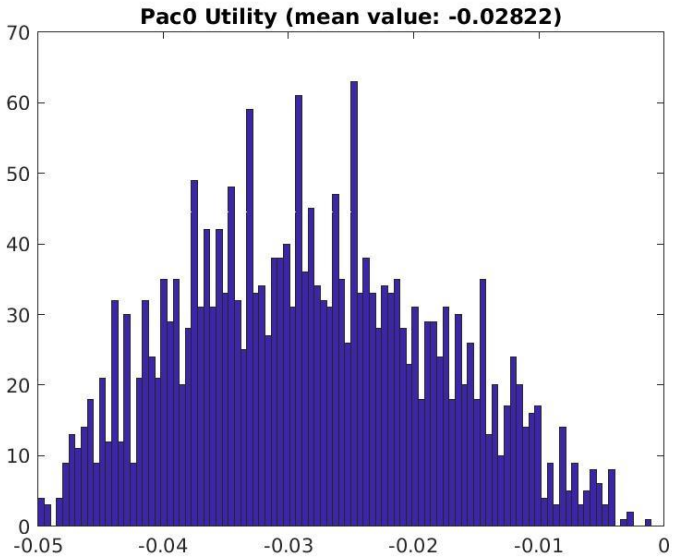

Figura 42. Utilidad para pacientes no sépticos para un mínimo de 48 horas en el conjunto de datos A.

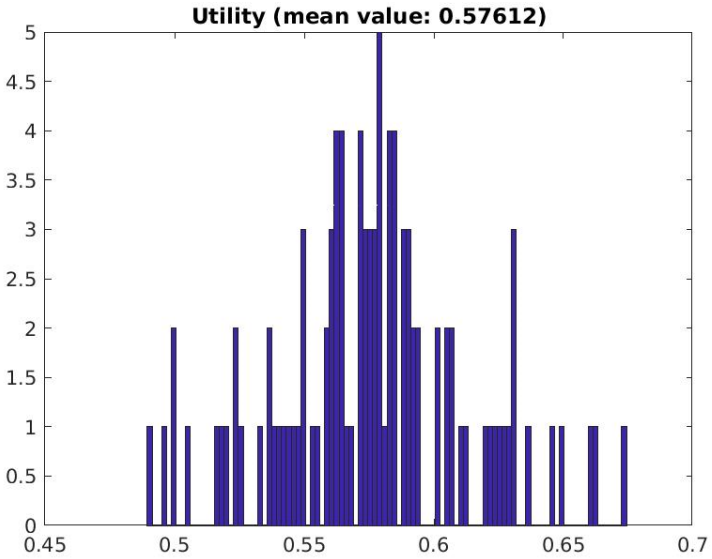

Figura 43. Utilidad total para un mínimo de 48 horas en el conjunto de datos A.

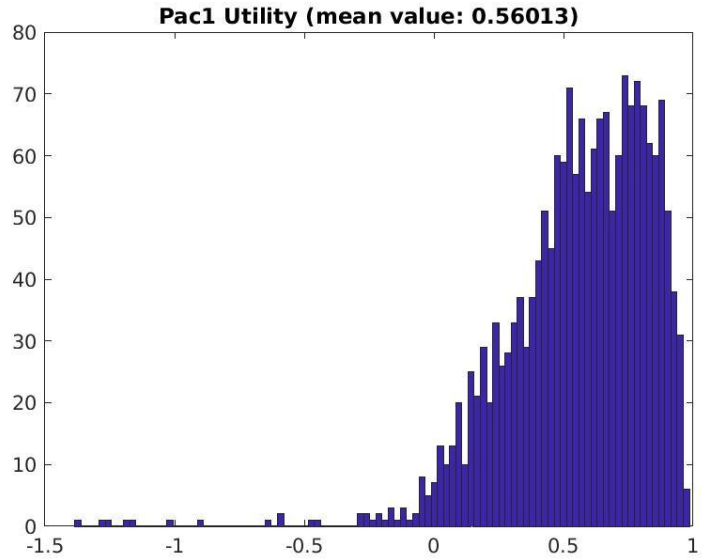

Figura 44. Utilidad para pacientes sépticos para un mínimo de 48 horas en el conjunto de datos B.

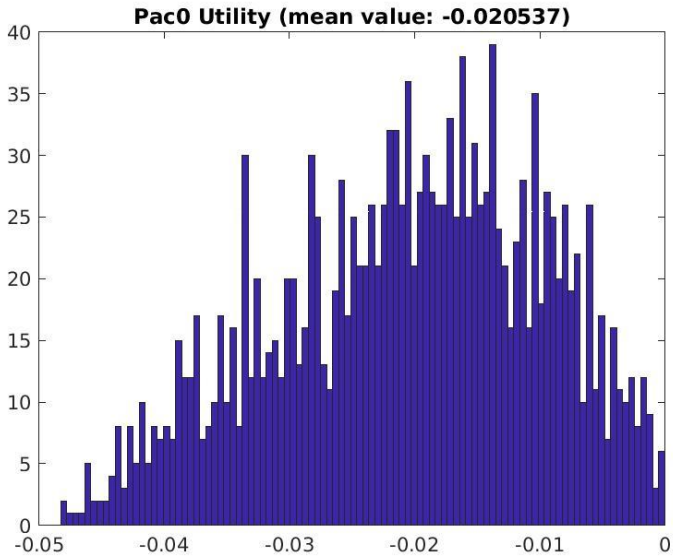

Figura 45. Utilidad para pacientes no sépticos para un mínimo de 48 horas en el conjunto de datos B.

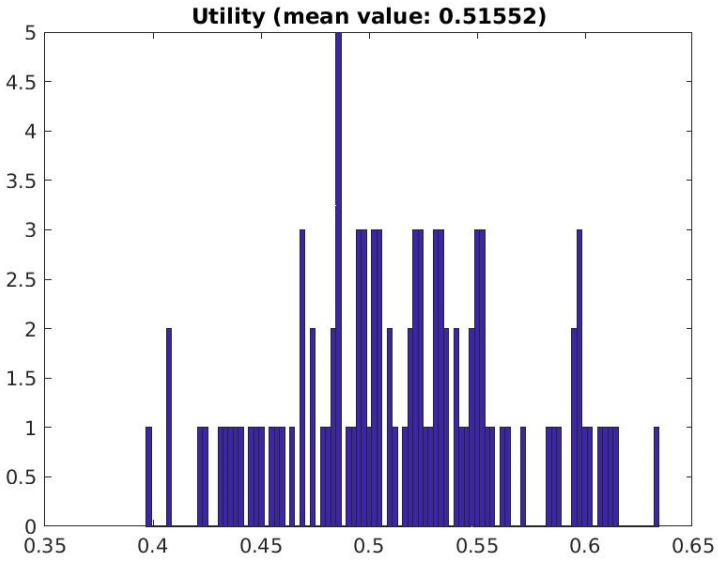

Figura 46. Utilidad total para un mínimo de 48 horas en el conjunto de datos B.

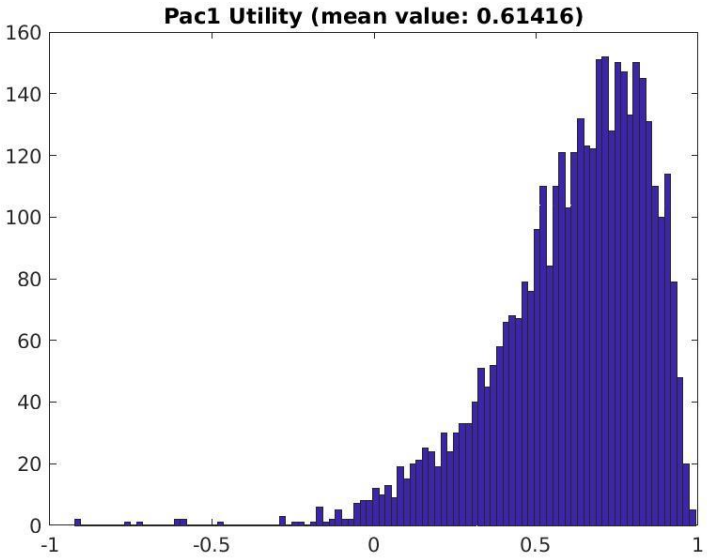

Figura 47. Utilidad para pacientes sépticos para un mínimo de 48 horas en el conjunto de datos A+B.

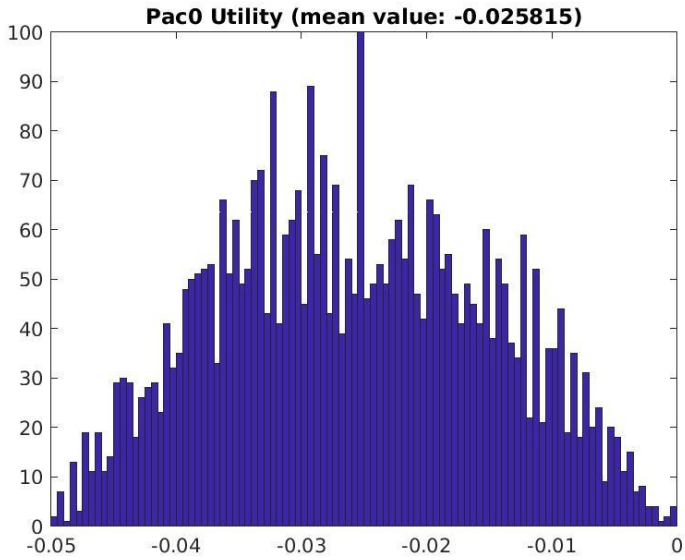

Figura 48. Utilidad para pacientes no sépticos para un mínimo de 48 horas en el conjunto de datos A+B.

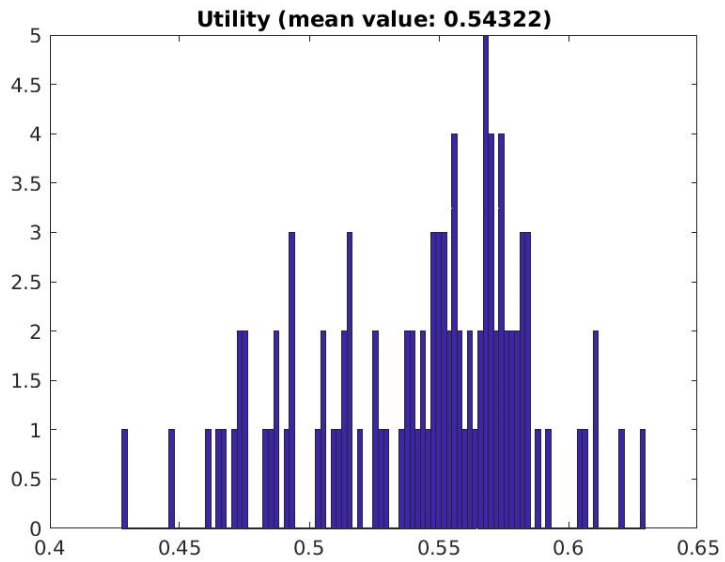

Figura 49. Utilidad total para un mínimo de 48 horas en el conjunto de datos A+B.

Anexo III: Utilidad para combinación de estrategias

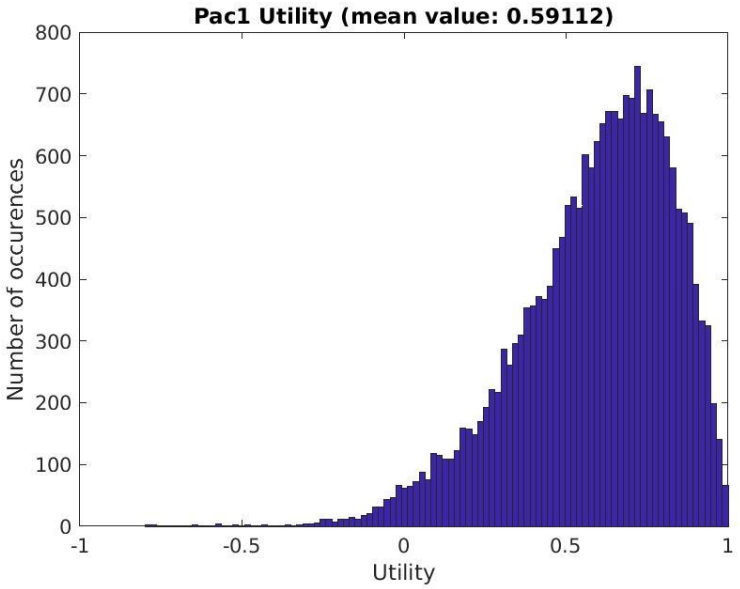

Figura 50. Utilidad para pacientes sépticos usando la combinación producto más majority voting para el conjunto de datos A.

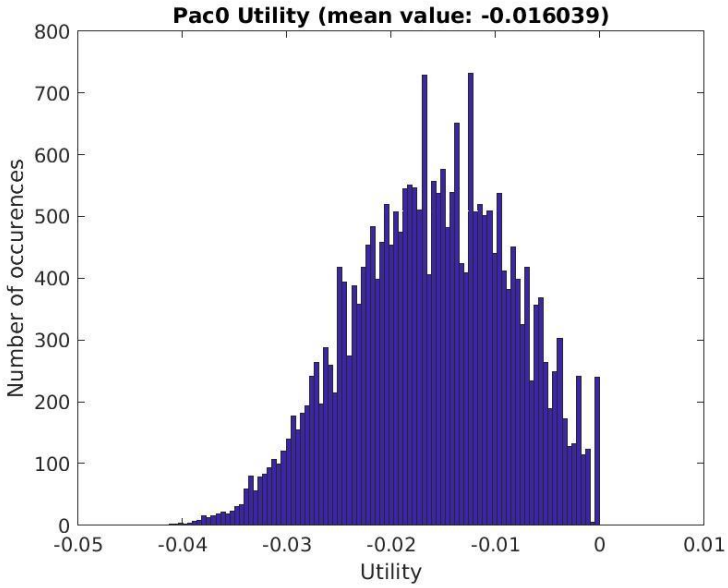

Figura 51. Utilidad para pacientes no sépticos usando la combinación producto más majority voting para el conjunto de datos A.

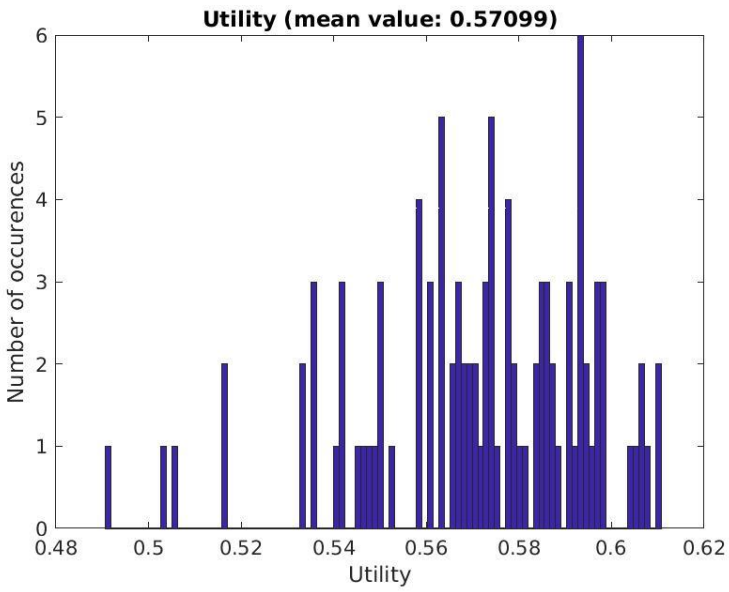

Figura 52. Utilidad total usando la combinación producto más majority voting para el conjunto de datos A.

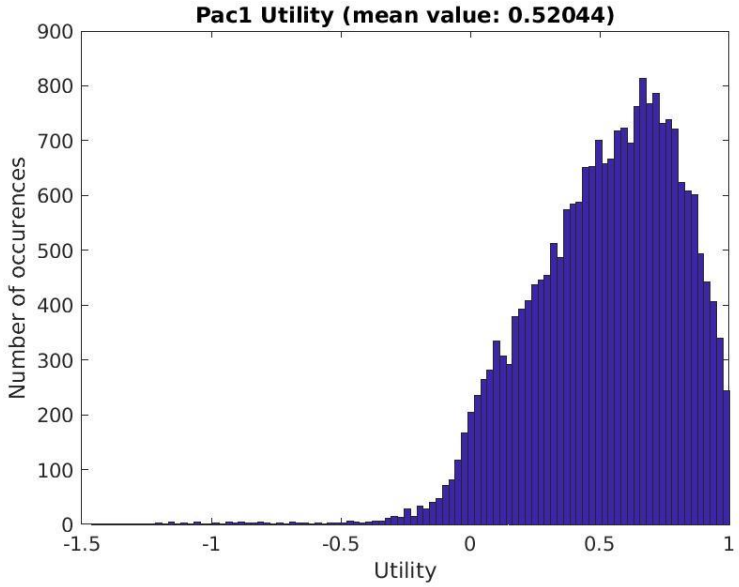

Figura 53. Utilidad para pacientes sépticos usando la combinación producto más majority voting para el conjunto de datos B.

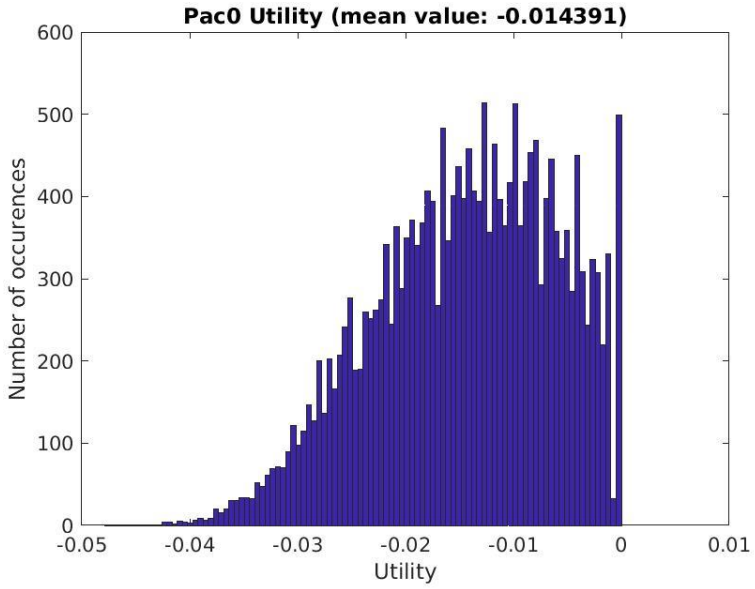

Figura 54. Utilidad para pacientes no sépticos usando la combinación producto más majority voting para el conjunto de datos B.

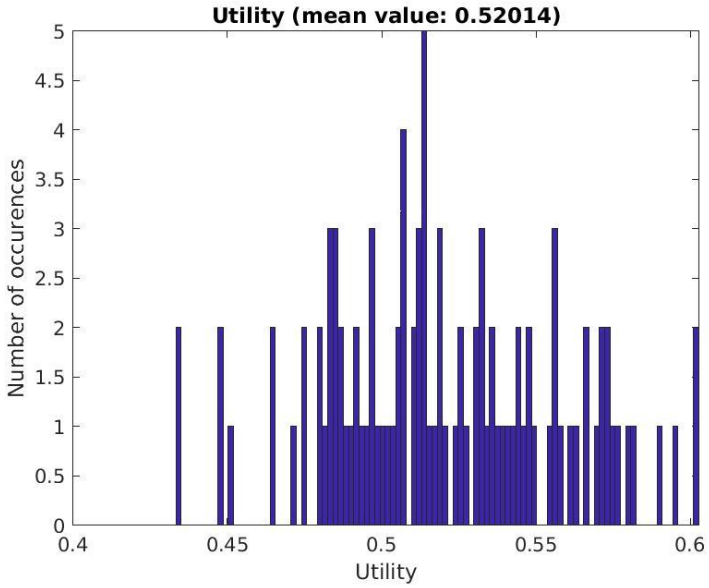

Figura 55. Utilidad total usando la combinación producto más majority voting para el conjunto de datos B.

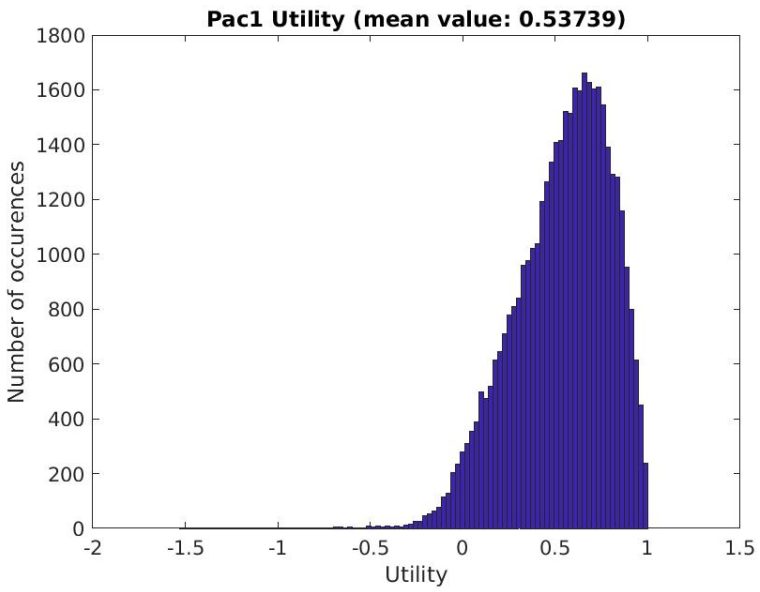

Figura 56. Utilidad para pacientes sépticos usando la combinación producto más majority voting para el conjunto de datos A+B.

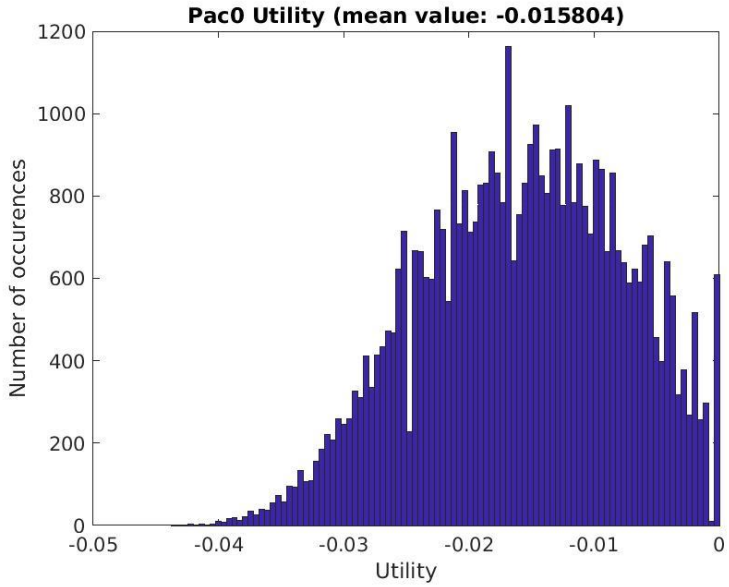

Figura 57. Utilidad para pacientes no sépticos usando la combinación producto más majority voting para el conjunto de datos A+B.

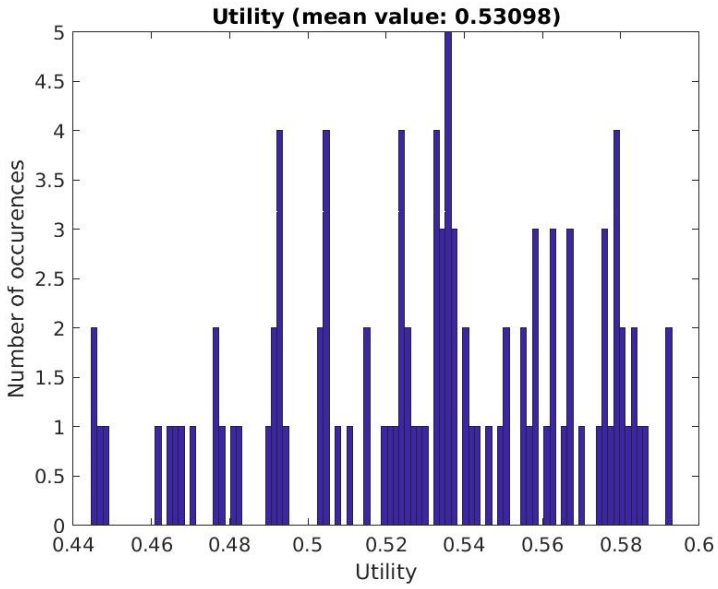

Figura 58. Utilidad total usando la combinación producto más majority voting para el conjunto de datos A+B.