MAT^2
MATerials MATemàtics
Volum 2021, treball no. 5, 36 pp. ISSN: 1887-1097
Publicació electrònica de divulgació del Departament de Matemàtiques
de la Universitat Autònoma de Barcelona
www.mat.uab.cat/web/matmat

Supervivencia en la UCI: un modelo de pronóstico usando conjuntos de clasificadores Bayesianos

Rosario Delgado

Prefacio

Este trabajo es una versión de carácter divulgativo del artículo [5] publicado recientemente, en colaboración con el Hospital de Mataró, y que ha tenido un cierto eco mediático. Las aplicaciones de la Inteligencia Artificial a la medicina como la que presentamos aquí, están de gran actualidad e interesan cada vez más. Y no es de extrañar, ya que todos somos potenciales usuarios de los servicios médicos y es lógico que nos interesemos por las mejoras que en ellos se puedan introducir.

1. Introducción

La medicina es en la actualidad una de las fronteras más interesantes del conocimiento humano. A nadie se le escapa que realizar un pronóstico preciso de un paciente (predecir de manera ajustada el curso de su enfermedad o patología) es clave en la atención médica de calidad. Aunque ya se ha empezado a trabajar en ello, el desarrollo de herramientas adecuadas a este propósito aún está poco avanzado, y la *Inteligencia Artificial* puede resultar muy útil para desarrollar metodologías que ayuden a los especialistas a realizar los pronósticos. Esto es especialmente necesario en las Unidades de Cuidados Intensivos (UCI) hospitalarias, donde la toma de decisiones clínicas sobre los pacientes críticos es un proceso costoso y complejo, que adolece de una excesiva variabilidad puesto que la opinión de los médicos especialistas depende en gran medida de su experiencia e instinto([9, 14]).

Aparte de la edad, las comorbilidades o los fallos orgánicos, existen otros aspectos relacionados con la muerte de los pacientes en la UCI, como puede ser una atención médica inadecuada, que también influyen en la duración de su estancia y los costes asociados, así como en la disminución de la calidad de vida de los pacientes al recibir el alta de la UCI ([6, 7, 13]). Para mejorar la calidad de la atención sanitaria, es importante establecer protocolos para la gestión del proceso asistencial. El enfoque tradicional para mejorar el rendimiento de las UCI se basa en el desarrollo de puntuaciones (escalas) para predecir la probabilidad de muerte de los pacientes ingresados, a partir de sus características. De ellas, el Acute Physiology And Chronic Health Evaluation (APACHE) en su versión II, es la escala más común para cuantificar la gravedad del paciente, y se basa en su evaluación durante las primeras 24 horas posteriores al ingreso en la UCI. La predicción de la probabilidad de muerte, según esta aproximación tradicional, se realiza mediante un modelo de regresión logística en el que APACHE II es uno de los regresores, validado en grupos previos de pacientes de UCI ([15]). Este enfoque, que explicaremos con un cierto detalle en la §4, presenta algunas limitaciones que lo hacen poco satisfactorio, tales como:

a) no incorpora variaciones entre unidades o regiones,

b) suele tener un buen comportamiento predictivo en poblaciones grandes, pero no en pequeñas,

c) es rígido, ya que si se desconoce el valor de alguna de las características del paciente (variables relacionadas con el tipo de ingreso y la puntuación APACHE), es incapaz de proporcionar una predicción,

d) en cierto sentido, ha quedado obsoleto, tras 40 años de aplicarse, teniendo en cuenta la evolución de la práctica médica en la actualidad. Además, se ha observado que el impacto de la edad en la supervivencia ha ido evolucionando, siendo uno de los ítems que más puntúa para APACHE, así como la esperanza de vida de los pacientes neoplásicos, coronarios o VIH, por ejemplo, realidad a la que el enfoque tradicional basado en APACHE no se ha adaptado.

Todo ello muestra la necesidad de desarrollar modelos fundamentados en nuevas metodologías para abordar estas deficiencias y mejorar las predicciones del riesgo de mortalidad para los pacientes de la UCI. La aplicación de la Inteligencia Artificial a la Medicina para construir modelos de pronóstico predictivo representa una oportunidad de mejora con respecto a los modelos basados en escalas.

En el artículo [5] se presenta una metodología de *Inteligencia Artificial* consistente en la construcción de un modelo predictivo de *Aprendizaje Automático*, que se valida con una base de datos real, viendo experimentalmente que da buenos resultados predictivos y evita los puntos débiles ya comentados

del enfoque tradicional, por lo que resulta una mejor alternativa. También se hace una revisión bibliográfica sobre trabajos en los que se han utilizado diferentes metodologías para ayudar a los expertos médicos en la toma de decisiones en las UCIs hospitalarias. El esquema de la Figura 1 ilustra la utilización del modelo predictivo que construimos, que se alimenta de la base de datos de pacientes de la UCI, a partir de la cual también se valida. Dado un nuevo caso de un paciente ingresado en la UCI, a partir de la evidencia dada por sus características, el modelo predictivo hará un pronóstico vital del paciente, prediciendo su probabilidad de vivir y de morir, y escogiendo como predicción lo más probable, así como su destino en el caso de que la predicción sea que vive, o la causa de la muerte, si la predicción es que muere.

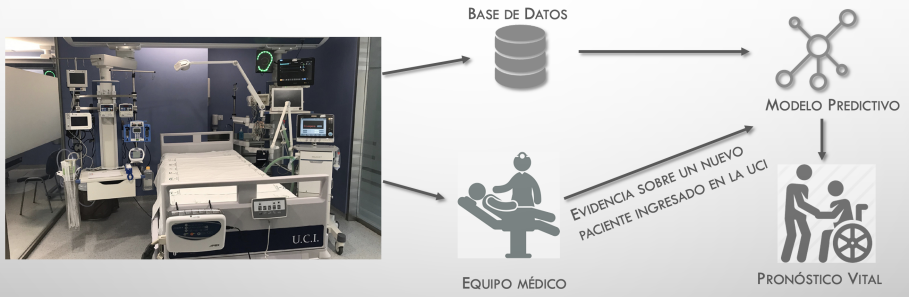


Figura 1: Esquema de funcionamiento del modelo predictivo para el pronóstico de pacientes críticos en la UCI.

2. Inteligencia Artificial, Aprendizaje Automático y Clasificadores

Inteligencia Artificial y *Aprendizaje Automático* son términos se usan a veces de manera intercambiable y con poca precisión, pero aunque están relacionados, no son lo mismo y es necesario explicarlos. Podemos decir que la **Inteligencia Artificial**, en general, engloba todos aquéllos algoritmos que intentan que una máquina imite las funciones cognitivas de los seres humanos, como el aprendizaje y la resolución de problemas.

El **Aprendizaje Automático** es el subcampo de la *Inteligencia Artificial* formado por aquéllos algoritmos que permiten analizar los datos, la construcción de modelos cuyo comportamiento mejora a medida que se nutren de más datos, y la toma de decisiones (predicciones). En otras palabras, son algoritmos que permiten a los modelos auto-aprenderse a partir de una base de datos, de tal manera que si ésta se actualiza, los modelos también, adaptándose a las modificaciones, y hacer predicciones informadas sobre los datos. El *Aprendizaje Automático* requiere matemáticas complejas y mucha programación, para implementar algoritmos que realicen diferentes tareas,

como *clustering*, regresión o clasificación. Nos centraremos en ésta última.

La **clasificación** consiste en lo siguiente: dado un conjunto de datos que consta de información sobre ciertos objetos o individuos (*instancias*, en general), relativa a algunas características suyas (*atributos*), y también a una variable categórica (la clase a la que pertenece cada instancia), con $r \geq 2$ clases posibles diferentes, digamos $y_1, \ldots, y_r$, **clasificar** una nueva instancia es inferir su clase, es decir, asignarle una clase, a partir de conocer sus características. El caso $r = 2$ se llama **binario** y será el que nos interesará principalmente, ya que la variable clase en nuestro caso será el *Resultado* de la estancia en la UCI del paciente, y tiene dos categorías posibles: `vivo` y `muerto`, según cómo abandone la UCI. En el caso binario, las clases suelen denominarse *positiva* + y *negativa* −, y lo usual es escoger como clase + la minoritaria, en este caso, `muerto` (los pacientes que mueren en la UCI representan menos del 15 %).

Un **clasificador** es un algoritmo de Aprendizaje Automático que permite realizar la tarea de clasificación. Concretamente, se dice que es de *Aprendizaje Automático* **Supervisado**, ya que en la fase de "entrenamiento" o "aprendizaje" del modelo, éste se construye a partir de una base de datos de casos resueltos en los que tanto los atributos como la clase se han consignado para las diferentes instancias (excepto valores faltantes), y se hace de manera que se procura optimizar un función objetivo. Después, en la fase de "validación" del modelo, se pueden comparar las predicciones de éste para la variable clase de las instancias de la base de datos que no se hayan utilizado para el aprendizaje, y los valores consignados para estas instancias, de tal manera que se puede saber si se ha acertado o no en la predicción y se pueden calcular diferentes métricas de comportamiento como medida de las discrepancias observadas. También los modelos de regresión son de *Aprendizaje Automático Supervisado*, pero en ese caso, la variable respuesta no es una variable clase categórica, como sucede con la clasificación, sino cuantitativa.

Hay diferentes metodologías del *Aprendizaje Automático Supervisado* que se pueden utilizar para la construcción de clasificadores. En particular, nos interesan los llamados **clasificadores probabilísticos**, que no sólo infieren o predicen la clase a la que pertenece una nueva instancia, sino que estiman la distribución de probabilidad sobre el conjunto de clases. Habitualmente, la clase inferida o predicha será la que tenga mayor probabilidad asociada, es decir, la clase a la que es más probable que pertenezca la instancia, dadas sus características (*evidencia*). Su probabilidad asociada se conoce como **nivel de confianza, CL** y esta manera de asignar la clase que predice el clasificador es el **criterio MAP** (por *Maximum A Posteriori*, ya que la probabilidad que asigna el modelo a cada clase dada la evidencia es una probabilidad *a posteriori*). Los clasificadores probabilísticos que siguen el criterio MAP, también conocido como **regla de decisión de Bayes**, reciben el nombre de **clasificadores Bayesianos**, y son los que consideramos en este trabajo.

El criterio MAP no es sólo intuitivo y razonable, sino que también cumple

una propiedad de optimalidad relacionada con la *función de pérdida* 0-1, que asigna un 0 a la tarea de clasificación si se acierta, y un 1 si se comete un error. Entonces, dada un evidencia E correspondiente a los valores de las variables características de una nueva instancia, si el clasificador asigna a esta instancia la clase y^* pero la clase realmente es y_j $(j = 1, \ldots, r)$, la pérdida asociada, que denotamos con la letra λ es:

$$\lambda(y_j,\, y^*) = \begin{cases} 0 & \text{si } y_j = y^* \\ 1 & \text{si } y_j \neq y^* \end{cases}$$

La propiedad de optimalidad es la siguiente:

Proposición 1 *La regla de decisión de Bayes (criterio MAP, que maximiza la probabilidad a posteriori)* **minimiza** *la pérdida esperada para la tarea de clasificación, si usamos la función de pérdida* 0-1.

Demostración: La pérdida debida a asignar a una instancia cuyas características son conocidas (evidencia E) la clase y^*, es una variable aleatoria que toma los valores $\lambda(y_j,\, y^*)$ $(j = 1, \ldots, r)$ con probabilidades $P(\text{clase} = y_j \,/\, E)$. Entonces, la pérdida esperada, que denotamos por R, será la esperanza matemática de esta variable aleatoria, esto es:

$$R(y^*) = \sum_{j=1,\ldots,r} \lambda(y_j,\, y^*)\, P(\text{clase} = y_j \,/\, E)\,.$$

¿Cuál es la clase que minimiza esta pérdida esperada, esto es, que minimiza R? Teniendo en cuenta que por la definición de λ, para una clase cualquiera y, $R(y)$ se puede re-escribir como:

$$\begin{aligned} R(y) &= \sum_{j=1,\ldots,r} \lambda(y_j,\, y)\, P(\text{clase} = y_j \,/\, E) = \sum_{j=1,\ldots,r,\, y_j \neq y} P(\text{clase} = y_j \,/\, E) \\ &= 1 - P(\text{clase} = y \,/\, E)\,, \end{aligned}$$

donde hemos usado que $\sum_{j=1,\ldots,r} P(\text{clase} = y_j \,/\, E) = 1$, vemos que la clase y que minimiza $R(y)$ coincide con la que maximiza $P(\text{clase} = y \,/\, E)$, que es la predicción según el criterio MAP. □

3. Conjuntos de clasificadores probabilísticos

Los clasificadores probabilísticos proporcionan predicciones que pueden ser útiles por sí mismas, pero también es posible agrupar clasificadores en conjuntos y combinar sus predicciones. Usar conjuntos de clasificadores (llamados *ensembles* en inglés) es una técnica que se basa en la idea de construir un nuevo clasificador a partir de la combinación de un conjunto de clasificadores, con el esperanza de mejorar su comportamiento predictivo ([1, 10, 12]).

Con este procedimiento perdemos la interpretabilidad de tener un solo clasificador, pero potencialmente ganamos en poder predictivo. Además, responde a una estrategia natural, ya que a menudo tendemos a consultar a diferentes personas antes de tomar nuestras decisiones importantes, y esto es especialmente cierto en el campo del diagnóstico clínico, donde no es infrecuente tener en cuenta las opiniones de varios expertos para llegar a la decisión final sobre un paciente.

Un tema de investigación importante en este campo es el de los **esquemas de combinación**, es decir, las reglas que se usan para combinar las predicciones de los clasificadores (véase [2]). En lugar de poner el énfasis en elegir un buen clasificador, lo ponemos en elegir un buen esquema de combinación, con la esperanza de que el mal comportamiento de algunos de los clasificadores del conjunto, sea compensado por el buen comportamiento de otros, dando como resultado un clasificador que se comporte bien y sea más robusto (Figura 2).

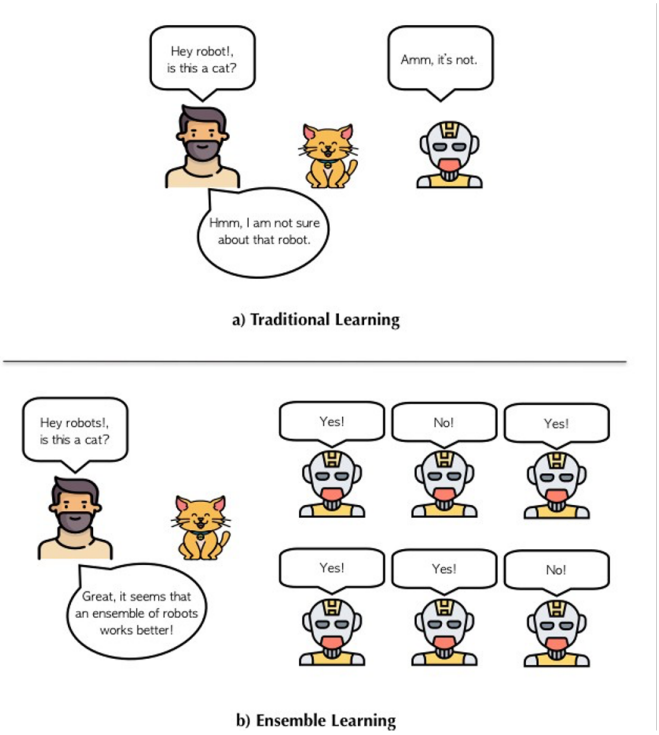


Figura 2: Ejemplo ilustrativo de un conjunto de clasificadores.

En general, inicialmente construimos $M \geq 2$ clasificadores probabilísticos diferentes, digamos $\mathcal{C}_1, \ldots, \mathcal{C}_M$, que formarán un conjunto de clasificadores cuya predicción será una combinación de sus predicciones. Como ilustración, supongamos que tenemos $M = 5$ clasificadores en el caso binario, de tal manera que para la variable *Resultado* de los pacientes en la UCI, los 3

primeros predicen `muerto` y el resto predicen `vivo`, siguiendo cada uno de ellos el criterio MAP, siendo que para los 3 primeros la probabilidad de `muerto` es 0.55, y para el resto es 0.10. La decisión final del conjunto se deriva usando una regla de combinación, que puede ser (véase [16], por ejemplo):

i) Una simple combinación de las clases predichas por cada clasificador $\mathcal{C}_1, \ldots, \mathcal{C}_5$, como la regla más conocida: la *mayoría de voto* simple (**MV**), que proporciona como predicción para el conjunto la mayoritaria para los clasificadores iniciales, en este caso `muerto`, como muestra la Tabla 1.

	Prob. de `muerto`	Predicción	Votos `muerto`	Votos `vivo`
$\mathcal{C}_1$	$p_1 = 0.55$	muerto	1	0
$\mathcal{C}_2$	$p_2 = 0.55$	muerto	1	0
$\mathcal{C}_3$	$p_3 = 0.55$	muerto	1	0
$\mathcal{C}_4$	$p_4 = 0.10$	vivo	0	1
$\mathcal{C}_5$	$p_5 = 0.10$	vivo	0	1
			3	2
MV		**muerto**		

Tabla 1: Ejemplo de uso de la regla de combinación *mayoría de voto* simple, MV.

ii) Una regla más compleja que combine las distribuciones de probabilidad predichas por cada clasificador $\mathcal{C}_1, \ldots, \mathcal{C}_5$, usando una función adecuada, como la media, el máximo, el mínimo o el producto (ver [10, 12]) y dé como predicción la clase que maximiza el valor de la función aplicada a las probabilidades predichas por cada clasificador base. Con el mismo ejemplo de la Tabla 1, podemos ver en la Tabla 2 que si usamos la media como función, lo que hacemos es calcular la media de las probabilidades de los clasificadores $\mathcal{C}_1, \ldots, \mathcal{C}_5$ para cada una de las clases, y escogemos como predicción la clase con mayor media. Denotamos por **EA** (por *Ensemble Average*) el conjunto de clasificadores construido con la función media.

	Prob. de `muerto`	Prob. de `vivo`	Predicción
$\mathcal{C}_1$	$p_1 = 0.55$	$1 - p_1 = 0.45$	muerto
$\mathcal{C}_2$	$p_2 = 0.55$	$1 - p_2 = 0.45$	muerto
$\mathcal{C}_3$	$p_3 = 0.55$	$1 - p_3 = 0.45$	muerto
$\mathcal{C}_4$	$p_4 = 0.10$	$1 - p_4 = 0.90$	vivo
$\mathcal{C}_5$	$p_5 = 0.10$	$1 - p_5 = 0.90$	vivo
EA	$\frac{1}{M}\sum_{i=1}^{M} p_i = 0.37$	$\frac{1}{M}\sum_{i=1}^{M}(1 - p_i) = \mathbf{0.63}$	**vivo**

Tabla 2: Ejemplo de uso de la regla de combinación *Ensemble Average*, EA.

iii) Una regla híbrida entre las dos situaciones anteriores (ver, por ejemplo, [2]).

En las anteriores reglas de combinación, todos los clasificadores del conjunto jugaban el mismo papel, pero no es infrecuente asignarles pesos distintos, según su importancia. Así, tendremos las versiones ponderadas de las reglas MV y EA, que denotaremos por **WMV** (*Weighted Majority Vote*) y **EWA** (*Ensemble Weighted Average*). Habitualmente, los pesos se determinan a través de un algoritmo de entrenamiento separado, generalmente en función de las precisiones estimadas de los clasificadores del conjunto (se dice entonces que las reglas de combinación son **entrenables**), aunque se podrían fijar con otros criterios. Siguiendo con el ejemplo anterior, si los pesos de los clasificadores $\mathcal{C}_1, \ldots, \mathcal{C}_5$ fuesen, respectivamente, $w_1 = 0.25$, $w_2 = 0.10$, $w_3 = 0.05$, $w_4 = 0.30$, $w_5 = 0.30$ (normalizados de manera que su suma sea 1), podemos ver que el clasificador WMV predice `vivo` ya que la suma de pesos de los clasificadores que predicen `muerto` es menor que la de los que predicen `vivo` (véase la Tabla 3).

	Pesos	Prob. de `muerto`	Predicción	Votos `muerto`	Votos `vivo`
$\mathcal{C}_1$	$w_1 = 0.25$	$p_1 = 0.55$	muerto	1	0
$\mathcal{C}_2$	$w_2 = 0.10$	$p_2 = 0.55$	muerto	1	0
$\mathcal{C}_3$	$w_3 = 0.05$	$p_3 = 0.55$	muerto	1	0
$\mathcal{C}_4$	$w_4 = 0.30$	$p_4 = 0.10$	vivo	0	1
$\mathcal{C}_5$	$w_5 = 0.30$	$p_5 = 0.10$	vivo	0	1
				$w_1 + w_2 + w_3 = 0.40$	$w_4 + w_5 = \mathbf{0.60}$
WMV			**vivo**		

Tabla 3: Ejemplo de uso de la regla de combinación *mayoría de voto* con pesos, WMV.

En cuanto al clasificador EWA, vemos en la Tabla 4 que predice la clase cuya suma ponderada de probabilidades es mayor, `vivo` en este caso.

	Pesos	Prob. de `muerto`	Prob. de `vivo`	Predicción
$\mathcal{C}_1$	$w_1 = 0.25$	$p_1 = 0.55$	$1 - p_1 = 0.45$	muerto
$\mathcal{C}_2$	$w_2 = 0.10$	$p_2 = 0.55$	$1 - p_2 = 0.45$	muerto
$\mathcal{C}_3$	$w_3 = 0.05$	$p_3 = 0.55$	$1 - p_3 = 0.45$	muerto
$\mathcal{C}_4$	$w_4 = 0.30$	$p_4 = 0.10$	$1 - p_4 = 0.90$	vivo
$\mathcal{C}_5$	$w_5 = 0.30$	$p_5 = 0.10$	$1 - p_5 = 0.90$	vivo
EWA		$\sum_{i=1}^{M} w_i\, p_i = 0.28$	$\sum_{i=1}^{M} w_i\,(1 - p_i) = \mathbf{0.72}$	**vivo**

Tabla 4: Ejemplo de uso de la regla de combinación *Ensemble Weighted Average*, EWA.

Obsérvese que la regla de combinación EA coincide con EWA si $w_1 = \ldots = w_M = \frac{1}{M}$, es decir, si todos los pesos son iguales, lo que equivale a no ponderar los clasificadores.

En [5] se compara la capacidad predictiva de un clasificador EWA (manera abreviada de referirnos a un conjunto de clasificadores que usa la regla de

combinación EWA; análogamente con el resto de reglas de combinación) con pesos adecuados, con los clasificadores WMV, con los mismos pesos, y también con los clasificadores EA y MV, construidos todos ellos como conjuntos de los mismos cinco clasificadores probabilísticos de tipo **red Bayesiana** (véase [3] y [4] para una introducción elemental a las redes Bayesianas). En cuanto a los pesos, se definen a partir de una transformación adecuada de una métrica de comportamiento, tal y como comentaremos en la Sección 5.

Al margen de la cuestión de la capacidad predictiva en sí, los conjuntos de clasificadores MV y WMV presentan un problema relacionado con el nivel de confianza asignado a sus predicciones. Según el criterio MAP, cada clasificador del conjunto tiene como nivel de confianza CL la probabilidad que le asigna a la predicción, que en el caso binario es > 0.5 (salvo empates). ¿Qué sucede con los conjuntos de clasificadores MV y WMV? ¿Podemos encontrar sendas fórmulas para el nivel de confianza que asignan a su predicción, que denotamos por $\mathrm{CL}_{\mathrm{MV}}$ y $\mathrm{CL}_{\mathrm{WMV}}$ respectivamente, en función de las probabilidades p_ℓ, $\ell = 1, \ldots, M$, siendo p_ℓ la probabilidad que asigna el clasificador ℓ-ésimo del conjunto a la clase escogida por el conjunto mediante la regla de combinación, que en el caso particular $M = 5$ son:

$$\begin{aligned}\mathrm{CL}_{\mathrm{MV}} = &\prod_{\ell=1}^{5} p_\ell + \sum_{j=1}^{5} \Big((1-p_j) \prod_{\substack{\ell=1 \\ \ell\neq j}}^{5} p_\ell \Big) \\ &+ \sum_{j=1}^{5} \sum_{\substack{k=1 \\ k>j}}^{5} \Big((1-p_j)\,(1-p_k) \prod_{\substack{\ell=1 \\ \ell\neq j,k}}^{5} p_\ell \Big), \\ \mathrm{CL}_{\mathrm{WMV}} = &\prod_{\ell=1}^{5} p_\ell + \sum_{r=1}^{4} \sum_{\substack{i_1,\ldots,i_r=1 \\ (i_1,\ldots,i_r)\in\Delta^r_{w_1,\ldots,w_5}}}^{5} \Big(\prod_{h=1}^{r} (1-p_{i_h}) \prod_{\substack{\ell=1 \\ \ell\neq i_1,\ldots,i_r}}^{5} p_\ell \Big),\end{aligned}$$

siendo

$$\Delta^r_{w_1,\ldots,w_5} = \Big\{ (i_1, \ldots, i_r) \,:\, 1 \leq i_1 < i_2 < \cdots < i_r \leq 5, \sum_{\ell=1}^{r} w_{i_\ell} < 0.5 \Big\}.$$

Notemos que para la regla WMV, el nivel de confianza se ha calculado como la probabilidad de que la suma de los pesos de los clasificadores del conjunto voten la clase que predice el conjunto sea > 0.5, lo que depende de los pesos (en el ejemplo no es posible tener empates, ya que no hay ningún subconjunto de los pesos cuya suma sea exactamente 0.5).

Si aplicamos la primera de las fórmulas al ejemplo de la Tabla 1, tenemos que la predicción dada por MV es `muerto`, con CL = $\mathbf{0.24731} < 0.5$,

mientras que con los pesos de la Tabla 3, la predicción dada por WMV es `vivo` con CL = **0.891** > 0.5, aplicando la segunda fórmula. Es decir, que podemos tener un nivel de confianza inferior a 0.5 en una predicción en el caso binario, cosa difícil de justificar desde un punto de vista intuitivo. Si cambiamos las probabilidades de los clasificadores del conjunto, manteniendo los pesos, podemos conseguir que el nivel de confianza del clasificador MV sea > 0.5 pero que el de WMV sea < 0.5. Un ejemplo es el de la Tabla 5. En cambio, tanto con el clasificador EA como con el EWA, el nivel de confianza es siempre > 0.5 por construcción: **0.63** en el primer caso (Tabla 2) y **0.72** en el segundo (Tabla 4).

	Pesos	Prob. de `muerto`	Predicción	Predicción MV	Predicción WMV
$\mathcal{C}_1$	$w_1 = 0.25$	$p_1 = 0.95$	muerto		
$\mathcal{C}_2$	$w_2 = 0.10$	$p_2 = 0.95$	muerto		
$\mathcal{C}_3$	$w_3 = 0.05$	$p_3 = 0.95$	muerto	muerto (**0.95324**)	vivo (**0.32725**)
$\mathcal{C}_4$	$w_4 = 0.30$	$p_4 = 0.45$	vivo		
$\mathcal{C}_5$	$w_5 = 0.30$	$p_5 = 0.45$	vivo		

Tabla 5: Ejemplo en el que $\text{CL}_{\text{MV}} > 0.5$ pero $\text{CL}_{\text{WMV}} < 0.5$. Entre paréntesis y en negrita, el nivel de confianza para cada conjunto de clasificadores.

4. Construcción del modelo

La base de datos a partir de la que construiremos el modelo corresponde a una cohorte de 2 510 pacientes críticos que estuvieron ingresados en la UCI del Hospital de Mataró en los años 2016 (661 pacientes), 2017 (693), 2018 (663) y 2019 (493). Con el objetivo de predecir la supervivencia/mortalidad en la UCI, en primer lugar, así como el destino de los pacientes al ser dados de alta de la UCI, o la causa de su muerte, según que se haya predicho `vivo` o `muerto`, se han considerado diferentes características (véase la Tabla 13, en el Apéndice A). Las UCIs polivalentes, como es nuestro caso, clasifican las características de los pacientes en cuatro categorías:

1. Características demográficas:

   Sexo (F_1)

   Edad (F_2)

2. Comorbilidades (índice de Charlson, F_3). Es una puntuación que se usa para evaluar la esperanza de vida a partir de la edad y otras características del paciente; cuanto mayor, peor es la esperanza de vida (más corta).

3. Ingreso

   Origen (procedencia, F_{17})
   Síndrome genérico (que causa el ingreso, F_{18})
   Septicemia (F_{19})
   Principal causa de ingreso (F_4 a F_{16})

4. Gravedad (en las primeras 24 horas del ingreso)

   Carga de trabajo UCI (requisitos terapéuticos, F_{20})
   APACHE II (F_{21})

Las variables Edad (F_2) y APACHE II (F_{21}) han sido discretizadas en 6 y 8 intervalos, respectivamente.

La aproximación habitual para evaluar el riesgo de muerte en la UCI se realiza a partir de la función de puntuación APACHE II y consiste en usar una *regresión logística* para estimar la probabilidad de muerte mediante

$$\frac{e^{logit}}{1+e^{logit}}$$

donde *logit* se obtiene a partir de la siguiente ecuación:

$$\begin{aligned} logit = -3.517 &+ 0.146 \times \text{APACHE II } (F_{21}) \\ &+ 0.603 \text{ (sólo si } F_{18} = \text{“Cirugía urgente”)} \\ &+ \text{coeficientes } \beta \text{ de } F_4 \text{ a } F_{15}, \text{ y de } F_{19}. \end{aligned} \tag{1}$$

(los coeficientes β son fijos y se pueden encontrar en la Tabla 14 del Apéndice A). Notemos que con esta aproximación tradicional, los datos se usarán sólo para validación, pero no para entrenar el modelo predictivo. Un primer intento de mejora de este procedimiento consiste en construir un modelo localmente recalibrado basado en la puntuación APACHE II pero aprendiendo los coeficientes de la ecuación (1) a partir de los datos, en vez de ser valores fijos. Llamamos a este modelo LR.APACHEII, y se construye también mediante *regresión logística*, a partir de los regresores: F_4 a F_{15}, F_{18} y F_{19}. El riesgo individual de muerte (probabilidad de `muerto` para la variable *Resultado*) con el modelo LR.APACHEII se calcula como

$$\frac{e^{\text{LR.}logit}}{1+e^{\text{LR.}logit}},$$

donde LR.*logit* se obtiene a partir de la siguiente ecuación, con coeficientes aprendidos a partir de los datos:

$$\begin{aligned} \text{LR.}logit = \alpha_0 &+ \alpha_1 \times \text{APACHE II } (F_{21}) + \alpha_2 \times F_{18} + \alpha_3 \times F_{19} \\ &+ \sum_{j=4}^{15} \alpha_j \times F_j. \end{aligned} \tag{2}$$

Los coeficientes α para el caso de aprenderlos a partir de toda la base de datos, se encuentran en la Tabla 15 en el Apéndice A, con los correspondientes p-valores para su significación estadística. Podemos observar que sólo hay un factor de protección, que es F_4 (en negrita), y cuáles son los factores de riesgo (el resto de regresores de la tabla).

Aunque vemos que el modelo LR.APACHEII mejora la aproximación tradicional, nuestro objetivo es construir un modelo de *Aprendizaje Automático Supervisado* que supere a ambos. En efecto, construimos un modelo jerárquico que constará de un primer conjunto de clasificadores para predecir la variable salida *Resultado* (`vivo`/`muerto`), y de dos conjuntos de clasificadores más, uno para predecir la variable salida *Destino* (al alta de la UCI), si la predicción ha sido `vivo`, y el otro para predecir la variable salida *Causa* (de la muerte) si ha sido `muerto`, como muestra el esquema de la Figura 3.

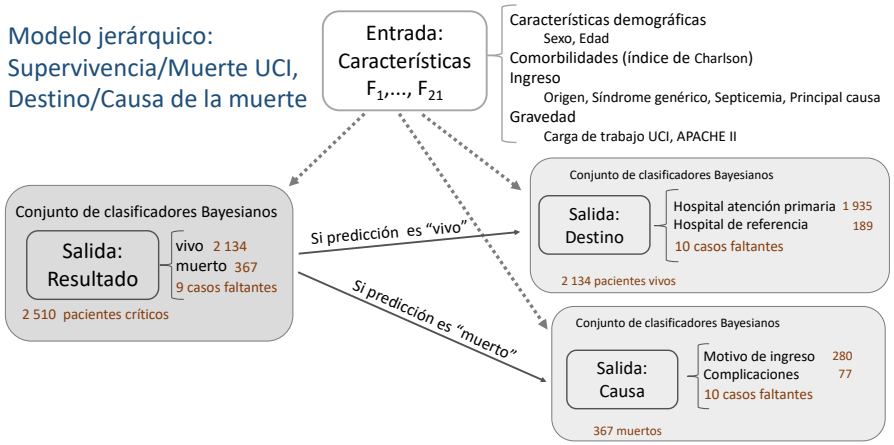


Figura 3: Esquema del modelo jerárquico para el pronóstico vital en la UCI.

Implementamos el meta-algoritmo para aprender los tres conjuntos de clasificadores y validarlos con el lenguaje de programación R. Construimos los conjuntos de clasificadores a partir de cinco clasificadores que son **redes Bayesianas (BN)**, y que aprendemos a patir de los datos. La ventaja de usar redes Bayesianas como clasificadores es que proporcionan una mayor comprensión de las relaciones entre las características de los individuos y su pronóstico vital, siendo esto una ventaja sobre otros métodos de aprendizaje automático supervisado que son *caja negra* ("*black box*"), como las redes neuronales.

Las BN son modelos probabilísticos de aprendizaje automático que representan las relaciones de dependencia entre las variables que afectan a un fenómeno de interés, en este caso, la supervivencia en la UCI. Dado un

conjunto de variables aleatorias discretas que serían las características del paciente (F_1–F_{21}) y la correspondiente variable salida clase (las variables *Resultado*, *Destino* o *Causa*, según corresponda), una BN es un modelo que consiste en una parte gráfica, que es un grafo acíclico dirigido DAG (por *Directed Acyclic Graph*), con nodos representando las variables, y arcos dirigidos entre ellas, de tal manera que no se permiten ciclos, esto es, caminos que siguiendo el sentido de los arcos salgan de un nodo y vuelven a él, y una distribución de probabilidad conjunta de las variables. Los arcos dirigidos entre los nodos representan dependencias condicionales (no necesariamente causales) entre las variables, gobernadas por la **condición de Markov**, que establece que cada nodo en el DAG es independiente de aquellos que no son sus descendientes (es decir, los nodos a los que no se puede acceder desde el nodo dado, siguiendo el sentido de los arcos dirigidos) dado que se conocen sus padres (que son los nodos origen de arcos dirigidos que finalizan en el nodo dado). Una red Bayesiana que se utiliza para clasificar casos en un conjunto de categorías o clases es un clasificador Bayesiano, ya que asigna la clase siguiendo el criterio MAP.

El aprendizaje de las redes Bayesianas a partir de la base de datos consta de dos partes: aprendizaje de la estructura (DAG), y aprendizaje de los parámetros, que son las probabilidades condicionadas de cada nodo en el DAG a sus padres (es equivalente conocer los parámetros a conocer la distribución de probabilidad conjunta de todas las variables, ya que ésta se obtiene de aquéllos por la **regla de la cadena**, como producto de las probabilidades condicionadas de cada nodo a sus padres). Para el aprendizaje (estimación) de los parámetros, seguiremos la aproximación de la Máxima Verosimilitud, que es un método general y de amplia aplicación en Estadística. Por lo que se refiere al aprendizaje de la estructura (DAG), se hará buscando heurísticamente un pseudo-máximo de una función de puntuación (*score*) que es básicamente el logaritmo de la función de verosimilitud, pero con un término de penalización por complejidad, que penaliza más cuanto mayor es el número de parámetros del modelo (nótese que el número de parámetros depende de la estructura). Trabajaremos con dos funciones de puntuación diferentes, que son las habituales:

$$\text{BIC (Bayesian Information Criterion)} = \log Lik - d\,\frac{\log M}{2}$$
$$\text{AIC (Akaike Information Criterion)} = \log Lik - d\,,$$

donde log indica logaritmo, Lik es la función de verosimilitud asociada al modelo y los datos de aprendizaje, M es el número de casos en el conjunto de datos de aprendizaje, y d es el índice de complejidad del modelo, definido como la dimensión del espacio de parámetros (número máximo de parámetros no redundantes). Como puede observarse en la definición, los términos de penalización por complejidad son $d\,\frac{\log M}{2}$ para BIC, y d para AIC, con lo que

AIC penaliza menos por complejidad que BIC (si $M > 7$), dando lugar a redes Bayesianas más conectadas, es decir, con más arcos dirigidos.

En el proceso de aprendizaje de la estructura se pueden imponer diferentes restricciones sobre los arcos dirigidos, obligando algunos de ellos (*whitelist*) o prohibiéndolos (*blacklist*), lo que da lugar a diferentes tipologías de redes Bayesianas. Concretamente, las que vamos a considerar están resumidas en la Tabla 6.

Red Bayesiana	*Score*	Restricción en los arcos dirigidos para el aprendizaje del DAG
$\mathcal{C}_1$ (Naive)		DAG fijo. Whitelist: de la clase a cada característica. Blacklist: entre características
$\mathcal{C}_2$ (Augmented Naive)	BIC	Whitelist: de la clase a cada característica
$\mathcal{C}_3$	AIC	Blacklist: de cada característica a la clase
$\mathcal{C}_4$ (Augmented Naive)	AIC	Whitelist: de la clase a cada característica
$\mathcal{C}_5$ (TAN)		Whitelist: de la clase a cada característica. Cada característica tiene un arco entrante adicional cuyo origen es otra característica

Tabla 6: Tipología de las cinco redes Bayesianas usadas para construir los conjuntos de clasificadores Bayesianos.

El Naive Bayes $\mathcal{C}_1$ tiene una estructura fija (DAG) que no se aprende a partir de los datos, con un arco dirigido de la variable clase a cada una de las características $\mathrm{F}_1, \ldots, \mathrm{F}_{21}$, y ninguno más (véase la Figura 4), y asume que las características son independientes entre sí dada la clase, por la **condición de Markov**, lo que puede ser poco realista en muchas aplicaciones, aunque incluso en esa circunstancia, el Naive Bayes ha demostrado ser un buen clasificador desde el punto de vista predictivo. Los otros cuatro clasificadores son diferentes intentos de mejorar la clasificación relajando este supuesto y tratando, al mismo tiempo, de mantener la simplicidad y la eficiencia tanto como sea posible. En particular, el TAN (Tree Augmented Naive) $\mathcal{C}_5$ relaja la suposición mediante una estructura de árbol, en la que cada característica depende únicamente de la variable clase y de otra característica. El resto de las redes Bayesianas que consideramos son de tipo Augmented Naive ($\mathcal{C}_2$ y $\mathcal{C}_4$, permitiendo arcos dirigidos adicionales a los del Naive Bayes entre las características), o redes Bayesianas de tipo clasificador, $\mathcal{C}_3$, sin restricciones adicionales, es decir, con la única restricción de no permitir arcos dirigidos de las características a la clase (este tipo de arcos tampoco están permitidos en las otras tipologías que hemos considerado, pero no era necesario imponer esta restricción ya que se deriva del hecho de no permitir ciclos en el DAG). Puede verse un ejemplo sencillo que ilustra las diferencias entre Naive Bayes, TAN y Augmented Naive en la Figura 4.

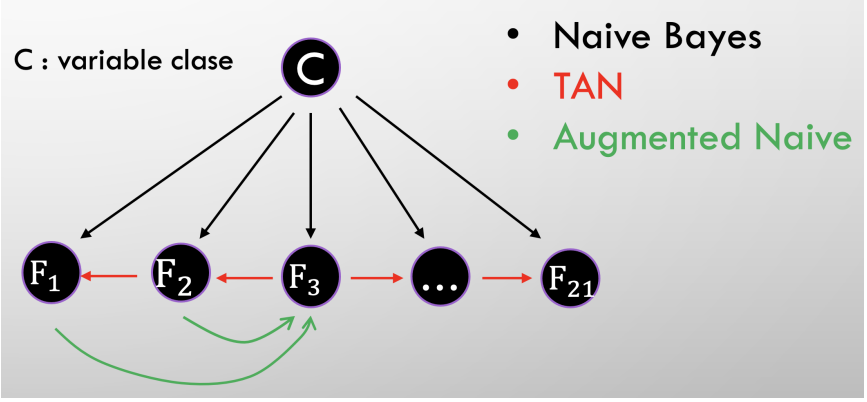


Figura 4: Tipologías de redes Bayesianas: Naive Bayes, Augmented Naive y TAN.

5. Validación y métricas de comportamiento

La aproximación más elemental al proceso de validación del modelo consiste en separar los casos de la base de datos original en dos partes disjuntas (aleatoriamente), usualmente en 80-20 % o 70-30 %, usando luego el subconjunto de casos mayor para construir (aprender) el modelo, y el resto como conjunto de validación para testar su capacidad predictiva. Esto se conoce como **split-validation**. Sin embargo, en el artículo hemos utilizado un método de validación un poco más sofisticado: el llamado k**-fold cross-validation**. Aunque k podría ser cualquier número natural, lo más habitual es utilizar $k = 10$ (excepto que la base de datos tenga pocos datos, caso en el que se utiliza una k menor). Empíricamente se ha visto que aumentar más la k no mejora el proceso de validación. Siguiendo este procedimiento de validación, los casos de la base de datos se dividen aleatoriamente en k partes (*folds*) aproximadamente del mismo tamaño, y se repite k veces lo siguiente: uno de los k *folds* (conteniendo aproximadamente el 10 % de los casos), uno diferente cada vez, se reserva como conjunto de validación, y su complementario (90 % de los casos) se usará como conjunto de entrenamiento a partir del cual se construirán los diferentes modelos predictivos que compararemos: tanto los 5 clasificadores base (redes Bayesianas $\mathcal{C}_1, \ldots, \mathcal{C}_5$), como los conjuntos de clasificadores con las diferentes reglas de combinación (MV, WMV, EA, EWA), y otros clasificadores de los habituales en Aprendizaje Automático. Estos clasificadores se utilizan para predecir la clase de los casos del conjunto de validación, para los que conocemos la clase que ha sido observada. De esta manera se podrán validar (y comparar) los clasificadores, atendiendo a su capacidad predictiva. En efecto, para cada clasificador y conjunto de validación, se genera una matriz de confusión binaria (en total, k matrices por clasificador) que resume los aciertos/errores cometidos, como la siguiente:

Matriz de confusión 2×2:

$$\begin{array}{cc} & \text{clase observada} \\ & \begin{array}{cc} + & - \end{array} \\ \text{clase predicha} & \begin{array}{c} + \\ - \end{array} \begin{pmatrix} t_p & f_p \\ f_n & t_n \end{pmatrix} \end{array}$$

Como es habitual, escogemos como clase positiva (+) la minoritaria y como clase negativa (−) la mayoritaria. En el caso de la variable clase *Resultado*, por ejemplo, la clase + es `muerto` y la clase − es `vivo`. Usamos la notación habitual:

t_p (*true positive*) es el número de casos del conjunto de validación que son de clase + y han sido correctamente clasificados por el clasificador,

t_n (*true negative*) es el número de casos del conjunto de validación que son de clase − y han sido correctamente clasificados por el clasificador,

f_p (*false positive*) es el número de casos del conjunto de validación que son de clase − y han sido erróneamente clasificados como +,

f_n (*false negative*) es el número de casos del conjunto de validación que son de clase + y han sido erróneamente clasificados como −,

n_+ es el número de casos del conjunto de validación que son de clase +, esto es, $n_+ = t_p + f_n$,

n_- es el número de casos del conjunto de validación que son de clase −, esto es, $n_- = f_p + t_n$,

N es el número total de casos en el conjunto de validación: $N = n_+ + n_- = t_p + f_n + f_p + t_n$.

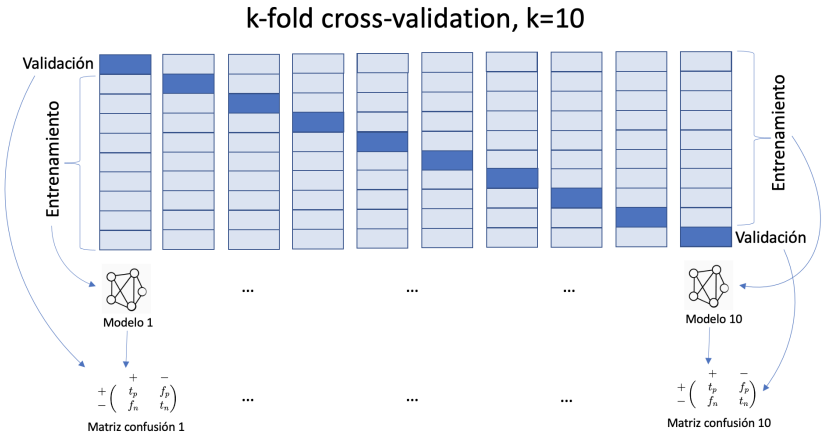


Figura 5: k-fold cross-validation con $k = 10$.

En la Figura 5 puede verse un esquema del procedimiento de validación que seguimos.

Repetimos el procedimiento 20 veces (*runs*) usando diferentes semillas aleatorias (*seeds*) para hacer la división de la base de datos en los $k = 10$ *folds*, y en cada *run* obtenemos $k = 10$ matrices de confusión para cada clasificador. De esta manera se evitan posibles sesgos en los resultados.

Dada una matriz de confusión binaria (2×2), podemos calcular diferentes métricas de comportamiento que cuantifican la capacidad predictiva del clasificador que la ha generado. El *accuracy* es la más habitual, y se define como la proporción de aciertos, esto es:

$$accuracy = \frac{t_p + t_n}{N} .$$

En nuestro caso, la base de datos es bastante desequilibrada, ya que la clase `muerto` no llega al 15 %. En este tipo de situaciones el *accuracy* no es una buena métrica de comportamiento, por la llamada **paradoja del accuracy**, que consiste en lo siguiente:

Paradoja del accuracy: Si hay desequilibrio (n_+ es mucho más pequeño que n_-), entonces un clasificador que siempre prediga $-$, acertará mucho (tendrá $accuracy = n_-/N$, que será grande), pero nunca predecirá bien ningún caso $+$.

Hay otras métricas de comportamiento de los clasificadores que tienen en cuenta no los aciertos en global, sino los de cada clase, y por tanto, lo bien que predice el clasificador no sólo la clase mayoritaria, sino también la minoritaria. Por ejemplo:

- **Precision** (Positive Predictive Rate): $t_p/(t_p+f_p)$ (proporción de aciertos dentro de los casos que clasificamos como $+$).
- **Recall** (Sensitivity o True Positive Rate): $t_p/(t_p + f_n)$ (proporción de aciertos dentro de los casos que realmente son $+$).
- **F-score** (media harmónica entre Precision y Recall): $2\,t_p/(2\,t_p + f_p + f_n)$.

Habitualmente, cuando se desea ponderar los clasificadores que forman un conjunto, los pesos se les asignan en función de su *accuracy*, que se ha de estimar con un procedimiento aparte. En este trabajo, en cambio, usaremos una métrica de comportamiento diferente del *accuracy*, que es función de las métricas Precision y Recall y, por tanto, es adecuada para trabajar con bases de datos desequilibradas: **AUPR** (Area Under the Precision-Recall curve), que es el área bajo la gráfica de la curva PR (gráfica de Precision como función de Recall), como se muestra en la Figura 6 (gráfica de la izquierda). Cuanto mayor es AUPR, mejor es el clasificador.

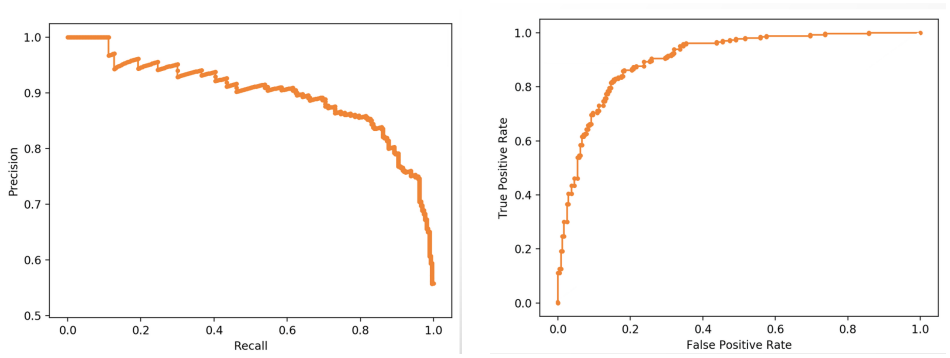


Figura 6: Ejemplo de curvas Precision-Recall, PR (izquierda) y Receiver Operating Characteristic, ROC (derecha).

Notemos que la métrica AUPR es similar a la más popular AUC (Area Under the ROC Curve), que es el área bajo la curva ROC (Receiver Operating Characteristic). La curva ROC es la gráfica de Recall como función del False Positive Rate, definido como $f_p/(f_p + t_n)$ (la proporción de errores de clasificación dentro de los casos que realmente son $-$), tal y como se muestra en la gráfica de la derecha de la Figura 6.

Teniendo esto en cuenta, asignaremos los pesos a los clasificadores $\mathcal{C}_1,\ldots,$ $\mathcal{C}_M$, del conjunto de clasificadores de la siguiente manera: la idea sería usar como pesos los valores estimados para cada clasificador de la métrica AUPR, que denotamos por A_i para el clasificador i-ésimo. Como $A_i \in [0,\, 1]$, resulta que $B_i = \frac{1}{2}\,(A_i + 1) \in [\frac{1}{2},\, 1]$ y, por tanto, para todo $i = 1, \ldots,\, M$,

$$\frac{B_i}{1 - B_i} = \frac{\frac{1}{2}\,(A_i + 1)}{1 - \frac{1}{2}\,(A_i + 1)} \geq 1\,,$$

así que podemos aplicar logaritmos y definir

$$h_i = \log\big(\frac{\frac{1}{2}\,(A_i + 1)}{1 - \frac{1}{2}\,(A_i + 1)}\big) \geq 0\,.$$

Finalmente, los pesos se obtienen normalizando para que la suma sea 1, de la siguiente manera:

$$w_i = \frac{h_i}{\sum_{j=1}^{M} h_j}\,.$$

Esta definición tiene sentido salvo que $h_i = 0$ para todos los clasificadores, lo que correspondería a que $A_i = 0$ para todo $i = 1, \ldots, M$, cosa que sólo sucedería en la situación extrema de que los M clasificadores realizasen la tarea de clasificación asignando las clases totalmente al azar, lo que no sucede en la práctica. La Tabla 7 muestra un ejemplo de cálculo de los pesos a partir

de los valores AUPR de los clasificadores con $M = 3$, y en la Figura 7 vemos una representación gráfica donde se observa una "dilatación", es decir, que con los pesos w_i magnificamos las diferencias en la relevancia de los clasificadores, medida a partir de la métrica AUPR.

Clasificador	AUPR A_i	$B_i = \frac{1}{2}(A_i + 1)$	$h_i = \log \frac{B_i}{1-B_i}$	$w_i = \frac{h_i}{\sum_{j=1}^{M} h_j}$
$\mathcal{C}_1$	0.25	0.625	0.5108256	0.2439739
$\mathcal{C}_2$	0.26	0.630	0.5322168	0.2541904
$\mathcal{C}_3$	0.49	0.745	1.072121	0.5120523

Tabla 7: Ejemplo de cálculo de los pesos a partir de la métrica AUPR.

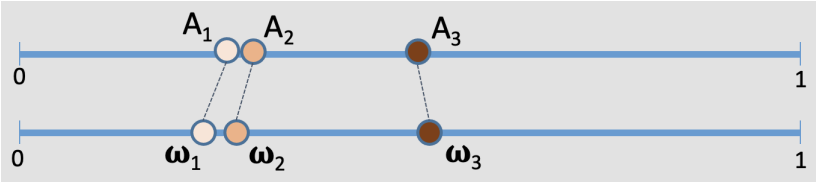


Figura 7: Valores de la métrica AUPR (A_i) y de sus correspondientes transformaciones en pesos para los clasificadores del ejemplo de la Tabla 7.

6. Resultados

Aplicando tests estadísticos adecuados a las métricas que obtenemos a partir de las matrices de confusión que se generan con las 20 repeticiones (*runs*) del procedimiento de k-fold cross-validation ($k = 10$), comparamos los conjuntos de clasificadores que construimos MV, WMV, EA y EWA, con los clasificadores del conjunto (las 5 redes Bayesianas $\mathcal{C}_1, \ldots, \mathcal{C}_5$), y entre ellos, así como con otros tipos de clasificadores de los usados en la actualidad de manera habitual (*neural networks NN, support vector machines SVM, random forests RF*). También los comparamos con la aproximación habitual para evaluar el riesgo de muerte en la UCI a partir de la función de puntuación APACHE II obtenida a partir de la fórmula (1), y con la versión localmente recalibrada que hemos introducido, LR.APACHEII, a partir de la fórmula (2).

La Tabla 8 nos muestra la media para las 20 repeticiones, de las medias y las desviaciones estándar, $\bar{x}$ y s respectivamente, sobre los 10 *folds*, para las métricas AUPR, F-score y AUC con la variable *Resultado*, y las denotamos por $\bar{\bar{x}}$ y $\bar{s}$, respectivamente. Más concretamente, si AUPR_i^j indica el valor

de la métrica AUPR obtenido con la repetición j para el clasificador que se valida con el *fold* i (y se aprende con su complementario), entonces para cada $j = 1, \ldots, 20$, introducimos las notaciones siguientes:

$$\bar{x}_j = \overline{\text{AUPR}}^j = \frac{1}{10}\sum_{i=1}^{10} \text{AUPR}_i^j\,, \quad s_j = \sqrt{\frac{1}{9}\sum_{i=1}^{10}\left(\text{AUPR}_i^j - \overline{\text{AUPR}}^j\right)}$$

y entonces,

$$\bar{\bar{x}} = \frac{1}{20}\sum_{j=1}^{20} \bar{x}_j\,, \quad \bar{s} = \frac{1}{20}\sum_{j=1}^{20} s_j$$

son los valores que se consignan en las columnas primera y segunda de la Tabla 8. Análogamente con las otras dos métricas. La métria AUC se incluye, aunque es menos adecuada que AUPR en una situación de desequilibrio entre las clases en la base de datos, como es nuestro caso, porque es muy popular en la literatura médica, para que se pueden comparar los resultados de nuestro estudio con otros realizados a partir de otras poblaciones y metodologías.

	AUPR		F-score		AUC	
	$\bar{\bar{x}}$	$\bar{s}$	$\bar{\bar{x}}$	$\bar{s}$	$\bar{\bar{x}}$	$\bar{s}$
$\mathcal{C}_1$	**0.52058 (5)**	0.08914	**0.54445 (1)**	0.06056	**0.87230 (3)**	0.02858
$\mathcal{C}_2$	0.50671	0.08514	**0.52763 (3)**	0.06642	0.85987	0.03192
$\mathcal{C}_3$	0.35161	0.13386	0.06805	0.02496	0.82825	0.04087
$\mathcal{C}_4$	0.48450	0.07944	0.49974	0.06621	0.83958	0.03605
$\mathcal{C}_5$	0.46424	0.07927	0.47569	0.06794	0.83277	0.03456
NN	0.27294	0.23893	0.43670	0.07689	0.70228	0.18079
SVM	0.43432	0.08309	0.32713	0.08628	0.79698	0.04014
RF	0.37071	0.08110	0.37567	0.08028	0.76864	0.04096
APACHEII	0.37899	0.09393			0.77518	0.04837
LR.APACHEII	0.42621	0.09342	0.30706	0.09075	0.83154	0.03744
MV	**0.52467 (3)**	0.08276	0.50274	0.06744	**0.86440 (4)**	0.03027
WMV	**0.52317 (4)**	0.08316	**0.51137 (5)**	0.06791	**0.86377 (5)**	0.03098
EA	**0.53829 (2)**	0.08510	**0.52354 (4)**	0.06666	**0.87913 (2)**	0.02538
EWA	**0.54131 (1)**	0.08423	**0.53270 (2)**	0.06766	**0.88026 (1)**	0.02522

Tabla 8: Media sobre las 20 repeticiones de las medias ($\bar{x}$) y las desviaciones estándar (s) obtenidas sobre los *folds*, respectivamente en las columnas $\bar{\bar{x}}$ y $\bar{s}$, para cada una de las tres métricas. En negrita los 5 clasificadores mejor posicionados para cada métrica. Las celdas en blanco indican que el F-score no se ha podido calcular por la disposición de ceros en las matrices de confusión.

Se observa una clara ventaja de los conjuntos de clasificadores, especialmente de EA y EWA, usando como métricas de comportamiento AUPR y AUC y también, aunque en menor medida, con F-score, respecto de los modelos predictivos basados en APACHE II que usan regresión logística, y respecto de los clasificadores habituales NN, SVM y RF, así como respecto de los clasificadores usados en los conjuntos, excepto $\mathcal{C}_1$, que es el Naive Bayes. Para ver si estas diferencias son significativas estadísticamente, haremos tests de hipótesis para cada una de las métricas.

A modo de ilustración exponemos algunos de los resultados. La Tabla 9 muestra el resultado de la comparativa únicamente entre los cuatro conjuntos de clasificadores; los valores consignados son los p-valores significativos (< 0.10) para los tests uni-laterales no-paramétricos de los rangos con signo de Wilcoxon, con hipótesis alternativa: "la mediana de los valores de la métrica AUPR para el conjunto de clasificadores correspondiente a la fila **es mayor** que la del correspondiente a la columna" realizados a partir de las muestras de tamaño $k = 10$, para cada repetición (*run*). Entre paréntesis se indica el número de las 20 repeticiones para las que hay significación estadística. Como es habitual, en esta y en el resto de tablas de resultados, $\cdot$ indica significación al 10 % (es decir, que el p-valor es < 0.10 y, por tanto, si el nivel de significación es $\alpha = 0.10$ o superior, se aceptaría la hipótesis alternativa), * al 5 % (ídem con $\alpha = 0.05$), ** al 1 % (ídem con $\alpha = 0.01$) y *** al 1‰ (ídem con $\alpha = 0.001$). Los p-valores de la Tabla 9 han sido ajustados usando el método de Holm-Bonferroni [8] para comparaciones múltiples (véase el Apéndice B para una breve explicación tanto de este método como del nivel de significación, del p-valor de un test estadístico y del método de Bonferroni).

Los p-valores de la Tabla 9 nos muestran que los conjuntos de clasificadores EWA y EA superan a WMV y MV con la métrica AUPR. Concretamente, EA supera a WMV en 8 repeticiones y a MV en 10, mientras que EWA lo hace en 8 y 15, respectivamente. Además, comparando los dos primeros entre sí, en 5 repeticiones ha habido diferencias significativas entre EWA y EA, y para todas ellas EWA ha sido significativamente mejor. Son estos resultados significativos? Para saber hasta qué punto lo son, calculamos los p-valores unilaterales para el test binomial exacto a favor de EA respecto de WMV y MV, y a favor de EWA respecto de EA, WMV y MV que son los de la Tabla 10. Por ejemplo, cuando comparamos EWA y EA, el p-valor exacto es

$$P(B(n = 5,\, p = 0.5) = 5) = 0.5^5 = 0.03125^*$$

y su interpretación es la siguiente: si realmente no hubiese diferencias significativas entre EWA y EA, la probabilidad de que, simplemente por azar, uno de los conjuntos apareciese como significativamente mejor que el otro, dado que se hubiesen observado diferencias significativas, sería de 0.5. La probabilidad de que simplemente por azar, en 5 de las 5 repeticiones para las que hay diferencias significativas entre ellos, EWA sea mejor que EA, es 0.03125. Como es pequeña la probabilidad de que esto se produzca simplemente por azar, obtenemos una significación en favor de que realmente el conjunto EWA es mejor (significativamente mayor mediana para AUPR) que el conjunto EA. Análogamente con el resto de comparaciones.

Podemos también comparar EWA y EA entre sí directamente, sin tener en cuenta ningún otro clasificador, con un test de Wilcoxon para muestras emparejadas. En ese caso, los correspondiente p-valores (sin ajustar) son los de la Tabla 11.

AUPR (Resultado)	EWA	EA	WMV	MV
EWA >		**(5)**	**(8)**	**(15)**
		0.0059**	0.029*	0.0342*
		0.029*	0.018*	0.098 ·
		0.0059**	0.024*	0.018*
		0.0342*	0.0117*	0.018*
		0.049*	0.039*	0.0098**
			0.029*	0.082 ·
			0.012*	0.012*
			0.055 ·	0.056 ·
				0.074 ·
				0.0059**
				0.029*
				0.093 ·
				0.059 ·
				0.012*
				0.018*
EA >			**(8)**	**(10)**
			0.049*	0.098 ·
			0.029*	0.027*
			0.056 ·	0.039*
			0.0645 ·	0.0146*
			0.024*	0.093 ·
			0.049*	0.068 ·
			0.012*	0.012*
			0.018*	0.034*
				0.029*
				0.021*

Tabla 9: p-valores ajustados por Holm-Bonferroni para la comparación entre los cuatro conjuntos de clasificadores para la variable *Resultado* y métrica de comportamiento AUPR. Entre paréntesis el número de repeticiones para los que el conjunto de clasificadores de la fila es mejor (mediana de AUPR significativamente mayor) que el de la columna. No hay más significaciones que las mostradas en la tabla.

AUPR (Resultado)	EWA	EA	WMV	MV
EWA >		$0.5^5 = 0.03125^*$	$0.5^8 = 0.00391^{**}$	$0.5^{15} = 3.05176 \times 10^{-5***}$
EA >			$0.5^8 = 0.00391^{**}$	$0.5^{10} = 0.00098^{***}$

Tabla 10: p-valores unilaterales del test binomial exacto a favor del clasificador de la fila versus el clasificador de la columna, calculados a partir de la información de la Tabla 9.

AUPR (Resultado)	run	**3**	**5**	**9**	14	**15**	18	**19**	20
EWA>EA	p-valor	0.00098***	0.014*	0.00098***	0.042*	0.0068**	0.08 ·	0.024*	0.042*

Tabla 11: p-valores (no ajustados) correspondientes a las comparaciones unilaterales entre EWA y EA en la Tabla 9, pero sólo considerando estos dos conjuntos de clasificadores. En negrita las 5 repeticiones correspondientes a los p-valores ajustados reportados en la Tabla 9.

Observemos que los p-valores no ajustados de la Tabla 11 son menores (más significativos) que los correspondientes ajustados de la Tabla 9, y ade-

más tenemos tres repeticiones adicionales para las que hay diferencias significativas, también a favor de EWA. El p-valor unilateral para el test binomial exacto a favor de EWA (respecto de EA, ahora sin ajustar los p-valores por comparaciones múltiples) es:

$$P(B(n = 8,\, p = 0.5) = 8) = 0.5^8 = \mathbf{0.00391}^{**}$$

y vemos que ha disminuido, mostrando mayor significación a favor de EWA.

No detallaremos el resto de los resultados, que son similares a los observados para AUPR, cuando utilizamos las métricas de comportamiento F-score y AUC, y que también muestran que EWA es estadísticamente significativamente mejor que que el resto de los clasificadores, así como que el clasificador LR.APACHEII es mejor que APACHEII, esto es, que es mejor aprender los parámetros de la recta de regresión de la base de datos, que utilizar unos valores prefijados.

7. Conclusiones

Es poco probable que el software inteligente reemplace al médico en el diagnóstico y pronóstico de sus pacientes, al menos a corto y medio plazo. Sin embargo, sí es probable que en un futuro próximo, y de hecho ya está empezando a pasar, los sistemas expertos de *Aprendizaje Automático* actúen como agentes inteligentes para problemas especializados y complicados, y ayuden a mejorar el desempeño del especialista. El objetivo principal de este trabajo ha sido demostrar la viabilidad y los beneficios de recopilar de forma rutinaria información de pacientes críticos ingresados en las UCI hospitalarias. Nuestra investigación está dirigida a la construcción de un clasificador jerárquico de *Aprendizaje Automático Supervisado* que permita predecir el riesgo de muerte en la UCI, así como el destino para quienes sobreviven a su estancia en la UCI, o la causa de la muerte para el resto, a partir de diferentes características del paciente.

En cada paso de esta clasificación jerárquica se utiliza un clasificador EWA, que es un conjunto de cinco redes bayesianas diferentes, con la regla de combinación del promedio ponderado con pesos apropiados, que se obtienen de las estimaciones de los valores de la métrica AUPR de los clasificadores usados para construir el conjunto (ya que como la base de datos es desequilibrada en cuanto a las dos categorías: `vivo/muerto`, es preferible esta métrica a la más habitual, el *accuracy*), mediante una transformación adecuada. Hasta donde sabemos, este es un enfoque novedoso para el pronóstico vital de pacientes en las UCI.

En la fase de validación, tanto con la métrica AUPR como con las otras dos métricas consideradas, F-score y AUC, hemos comparado el rendimiento predictivo de los clasificadores y hemos visto que es mejor el de EWA que el de sus contrincantes, que son:

- los clasificadores bayesianos a partir de los cuales se ha construido, $\mathcal{C}_1, \ldots, \mathcal{C}_5$,
- otros métodos habituales de *Aprendizaje Automático* de última generación (neural networks, support vector machine, random forests),
- conjuntos de clasificadores obtenidos mediante otras reglas de combinación (MV, WMV, EA),
- modelos predictivos basados en la escala APACHE II mediante regresión logística. El enfoque tradicional de este tipo de metodología, que usa parámetros prefijados (modelo APACHEII) adolece de no incorporar elementos que la práctica clínica revelan como de gran valor, como el origen del paciente (variable F_{17} en nuestro modelo), ya que dependiendo del origen, el paciente puede presentar mayor o menor grado de fragilidad y, por tanto, tener una evolución muy diferente en la UCI. También hemos considerado una mejora de este modelo que es una recalibración local obtenida a partir de la base de datos (y que se comporta mejor que él), el modelo LR.APACHEII.

También profundizamos en la interpretabilidad del conjunto de clasificadores EWA, elegido como el mejor de los probados, utilizando diferentes técnicas. En particular, usamos las tablas de probabilidad condicionada de la variable *Resultado* a las características $F_1, \ldots, F_{21}$, estimadas a partir del modelo EWA aprendido a partir de toda la base de datos, para ordenar las características atendiendo a su "fuerza" o importancia para la predicción de la variable *Resultado*. Las cinco más importante, ordenadas de más a menos, se encuentran en la Tabla 12, con las categorías que maximizan el riesgo de muerte, obtenidas a partir de las tablas del Apéndice C.

Característica	Categoría que maximiza el riesgo de muerte	Riesgo de muerte
F_{21}: APACHE II	> 34	64 %
F_{10}: CRA (Cardio Respiratory Arrest)	sí	62 %
F_{20}: Carga de trabajo UCI	Inestable en coma o conmoción	44 %
F_{17}: Origen	Emergencias extra-hospitalarias	46 %
F_{18}: Síndrome genérico	Médico	19.5 %

Tabla 12: Las 5 características con más fuerza para la predicción del riesgo de muerte según el modelo EWA, con la categoría que maximiza el riesgo, y su riesgo asociado.

La siguiente característica en importancia es F_4: ACS (Acute Coronary Syndrome), la única de entre la categoría "Principal causa de ingreso" que actúa como factor de protección, es decir, tal que que su presencia reduce el riesgo de muerte. A partir de la Tabla 18 (Apéndice C) obtenida con el modelo EWA podemos estimar, por ejemplo el **Odds Ratio** (OR) en favor de `muerto` para la presencia de ACS ($F_4 =$ sí) respecto de su ausencia ($F_4 =$ no), independientemente de las otras características:

$$\mathrm{OR}_{F_4=\text{sí}\,/\,F_4=\text{no}} = \frac{0.00390/(1-0.00390)}{0.11757/(1-0.11757)} = 0.02939$$

Como $1/0.02939 = 34.02518$, resulta que los *odds* en favor de `muerto` se dividen aproximadamente por 34 cuando se presenta ACS respecto de cuando no. Este hecho parece contraintuitivo, pero en realidad no lo es. De hecho, la práctica clínica indica que entre los pacientes ingresados en la UCI, aquéllos cuyo "Síndrome genérico" F_{18} = Coronario, claramente asociado con ACS (F_4 = sí), son los que tienen mejor pronóstico, ya que si no es ésta la principal causa de ingreso, será otra de mayor gravedad. Por ejemplo, si un paciente presenta fallo respiratorio (RF, F_5 = sí), que está asociado con "Síndrome genérico" F_{18} = Médico, su pronóstico empeora (se incrementa la predicción de su riesgo de muerte), en consonancia con la información de la Tabla 12. Téngase en cuenta que el 95.2 % de los pacientes con F_{18} = "Coronario" presenta ACS, mientras que sólo el 0.7 % de ellos presenta RF. Por otro lado, de los pacientes con F_{18} = "Médico", el 47.8 % presenta RF pero únicamente el 2.8 % presenta ACS.

De las 5 características más importante desde el punto de vista predictivo que aparecen en la Tabla 12, sólo una es de la categoría "Principal causa de ingreso", F_{10} (CRA), un factor de riesgo importante. A partir de la Tabla 20 podemos calcular el OR en favor de `muerto` correspondiendo a la presencia de CRA (F_{10} = sí) respecto de su ausencia (F_{10} = no), que es:

$$\mathrm{OR}_{F_{10}=\text{sí}\,/\,F_{10}=\text{no}} = \frac{0.62298/(1-0.62298)}{0.12604/(1-0.12604)} = 11.45758$$

Esto es, los *odds* a favor de `muerte` se multiplican por aproximadamente 11.5 cuando el paciente presenta CRA, respecto de cuando no es así. Es importante resaltar que estos resultados son generales, pero que para un paciente específico, conociendo algunas de sus características (por ejemplo sexo o edad), estos resultados pueden ajustarse para obtener una evaluación más precisa del riesgo de muerte mediante el cálculo del OR.

Sería muy interesante extrapolar nuestro modelo a una base de datos compuesta por casos de diferentes UCI, lo que permitiría comparar el rendimiento de las diferentes unidades; para ello se introducirían en el modelo las características de un paciente típico y se predeciría el riesgo de muerte en cada UCI. Esta herramienta también permitiría realizar un estudio longitudinal y analizar la mejora en el tiempo de los procesos asistenciales de una UCI específica, así como adecuar el modelo a los diferentes tipos de UCI, desde el centro de trauma hasta la especializada en problemas respiratorios o cardiovasculares.

En la medida en que pueda ayudar a los médicos a tomar decisiones terapéuticas adaptadas al paciente, y a las autoridades sanitarias a gestionar de forma más óptima los recursos disponibles, siempre escasos, la metodología del *Aprendizaje Automático* en general, que se basa en el aprendizaje de los modelos a partir de la información recogida en una base de datos adecuada para poder hacer predicciones, y en particular la que presentamos en este trabajo para estimar el riesgo de muerte y predecir el destino al alta de la

UCI o la causa de la muerte, es una herramienta muy útil y prometedora, con una importante aplicabilidad clínica. Esta es, en resumen, la contribución de nuestro trabajo.

Apéndice A: Tablas adicionales

1. Características demográficas	% respecto de valores no faltantes
F_1 : Sexo	
Hombre	63.6 %
Mujer	36.4 %
F_2 : Edad	Mediana: 70
	Q_1, Q_3 : 59, 79
Rangos:	
< 45	8.8 %
45–54	9.8 %
55–64	19.4 %
65–74	24.9 %
75–84	26.2 %
> 84	10.9 %
2. Comorbilidades	
F_3 : índice de Charlson	
0	31.7 %
1	24.2 %
2	15.9 %
3	10.5 %
> 3	17.7 %
3. Ingreso	
F_{17} : Origen (procedencia)	
Planta	20.2 %
Quirófano	14.0 %
Emergencias	41.0 %
Emergencias extra-hospitalarias	1.7 %
Otro hospital	23.1 %
F_{18} : Síndrome genérico (que causa el ingreso)	
Cirugía electiva	6.5 %
Cirugía urgente	9.8 %
Coronario	17.5 %
Médico	64.3 %
Trauma	1.9 %
F_{19} : Septicemia	
sí	35.7 %
no	64.3 %
Principal causa de ingreso (sí/no)	% de **sí**
F_4: ACS (Acute Coronary Syndrome)	18.7 %

F_5: RF (Respiratory Failure)	33.0 %
F_6: Shock	27.1 %
F_7: Coma	7.3 %
F_8: Renal F (Renal Failure)	4.1 %
F_9: Hepatic F (Hepatic Failure)	0.2 %
F_{10}: CRA (Cardio Respiratory Arrest)	4.8 %
F_{11}: ES (Elective Surgical)	6.7 %
F_{12}: Arrhythmia	4.1 %
F_{13}: CT (Cranial Trauma)	0.2 %
F_{14}: OT (Other Trauma)	1.3 %
F_{15}: Intoxication	1.0 %
F_{16}: Other syndromes	6.3 %
4. Gravedad (en las primeras 24 horas del ingreso)	
F_{20} : Carga de trabajo UCI (requisitos terapéuticos)	
Seguimiento médico	25.4 %
Inestable en coma o conmoción	22.4 %
Inestable sin coma ni conmoción	21.2 %
Seguimiento post-operatorio	5.1 %
Inestable post-operatorio	25.9 %
F_{21} : APACHE II	Mediana: 13 Q_1, Q_3 : 8, 18.25
Rangos:	
< 5	9.0 %
5–9	25.6 %
10–14	23.6 %
15–19	19.6 %
20–24	11.5 %
25–29	6.2 %
30–34	2.7 %
> 34	1.8 %
Variables Salida	
Resultado	
vivo	85.3 %
muerto	14.7 %
Destino (al alta de la UCI)	
Hospital de atención primaria	77.7 %
Hospital de referencia	7.6 %
Morgue	14.7 %
Causa (de la muerte)	
Motivo de ingreso	11.2 %
Complicaciones	3.1 %
No muerto	85.7 %

Tabla 13: Lista de variables del paciente. *Destino* = "Morgue" si *Resultado* = "muerto". *Causa* = "No muerto" si *Resultado* = "vivo". Se han fusionado las clases "Complicaciones sépticas" y "Complicaciones no-sépticas" (1.57 % y 1.53 %, respectivamente) para la variable *Causa*, en una única clase "Complicaciones".

Variables	Septicemia o Síndrome genérico no quirúrgico	No septicemia y Síndrome genérico quirúrgico
F_4	-0.191	-0.797
F_5	-0.890	-0.610
F_6	0.493	-0.797
F_7	-0.759	-1.150
F_8	-0.885	-0.196
F_9	0.501	-0.613
F_{10}	0.393	0.393
F_{11}		-0.248
F_{12}	-1.368	-0.797
F_{13}	-0.517	-0.955
F_{14}	-1.228	-1.684
F_{15}	-0.142	-0.196
F_{19}	0.113	

Tabla 14: Coeficientes β para la regresión logística tradicional usada para predecir la probabilidad de muerte a partir de APACHE II. "Síndrome genérico quirúrgico" significa F_{18} = Cirugía electiva o urgente, mientras que "Síndrome genérico no quirúrgico" es el complementario. Tabla adaptada de [11].

Variables	α estimada	p-valor	interpretación
Intercept	$\alpha_0 = -4.85711$	$5.72 \times 10^{-14***}$	0.00777[a]
APACHE II	$\alpha_1 = 0.11544$	$< 2 \times 10^{-16***}$	12.2 %[b]
F_{19}	$\alpha_3 = 0.48604$	0.00212^{**}	62.6 %[c]
F_4	$\alpha_4 = -1.83024$	0.03768^{*}	**84.1 %**[d]
F_5	$\alpha_5 = 0.62866$	0.00126^{**}	87.5 %[c]
F_6	$\alpha_6 = 0.52522$	0.00896^{**}	69.1 %[c]
F_7	$\alpha_7 = 0.49130$	$0.05651\,\cdot$	63.4 %[c]
F_{10}	$\alpha_{10} = 1.82376$	$4.38 \times 10^{-9***}$	519.5 %[c]

Tabla 15: Coeficientes α para la ecuación (2) (sólo aquéllos con p-valores significativos, esto es, > 0.10).(a): *odds* a favor de "muerto" cuando los regresores están en sus valores de referencia (todos iguales a "no", excepto APACHE II, que toma el valor "0"). (b): incremento en los *odds* a favor de "muerto" para un incremento unitario en APACHE II, sin modificar el resto de regresores. (c): incremento en los *odds* a favor de "muerto" cuando el regresor toma el valor "sí", con respecto a cuando toma el valor "no", sin modificar el resto de regresores. (d): decrecimiento en los *odds* en favor de "muerto" cuando F_4 toma el valor "sí", con respecto a cuando toma el valor "no", sin modificar el resto de regresores.

Apéndice B: el método de Holm-Bonferroni

En este apéndice se explica brevemente el método de Holm-Bonferroni de comparaciones múltiples. El problema de las comparaciones múltiples se presenta cuando en vez de una única hipótesis estadística, se desea constrastar varias a la vez. La situación típica es aquélla en la que se quieren hacer tests de comparaciones para parejas de elementos de un conjunto. Imaginemos que

tenemos un conjunto con r elementos (por ejemplo, diferentes clasificadores) y planteamos todos los posibles tests de comparación entre parejas, que son un total de $m = \binom{r}{2}$ (en nuestro caso $r = 4$ cuando se trata de comparar los conjuntos de clasificadores MV, WMV, EA y EWA, y $m = \binom{4}{2} = 6$ son las comparaciones múltiples entre parejas de clasificadores de estos 4). Para cada uno de estos tests de comparación entre parejas, digamos el test $i = 1, \ldots, m$, la hipótesis nula a constrastar es la de la igualdad, mientras que la alternativa, que denotamos por H_i, es la de que son diferentes (en el caso bilateral; en el uni-lateral sería análogo). Así que tenemos las hipótesis alternativas $H_1, \ldots, H_m$, y para cada una de ellas, el correspondiente p-valor[1] , digamos $p_1, \ldots, p_m$, que podemos suponer sin pérdida de generalidad ordenados de menor a mayor: $p_1 \leq p_2 \leq \cdots \leq p_m$, reordenando las hipótesis si es necesario.

El problema que se genera al realizar estas comparaciones de manera conjunta (*comparaciones múltiples*), cada una de ellas con nivel de significación α, es que si no hacemos ningún tipo de corrección o ajuste, la probabilidad de cometer el error de **aceptar alguna de las hipótesis H_i siendo falsa** (se conoce como *family-wise error*) no tiene por qué ser menor que α. De hecho, podría ser mucho mayor! Si denotemos por Δ el subconjunto de índices correspondientes a las hipótesis que realmente son falsas (desconocido, obviamente), esto es:

$$\Delta = \{j = 1, \ldots, m \,:\, H_j \text{ es falsa}\} \tag{3}$$

(asumimos que Δ no es el conjunto vacío, porque es caso contrario, el *family-wise error* sería imposible), denotamos su cardinal por δ (y, por tanto, asumimos que $\delta \in \{1, \ldots, m\}$, también desconocido), y tenemos que se cumple $P(\text{aceptar } H_j \,/\, H_j \text{ es falsa}) = p_j$, aceptando H_j si $p_j < \alpha$. Entonces, la probabilidad del *family-wise error* es:

$$\begin{aligned} P(\text{aceptar alguna } H_j \text{ con } j \in \Delta) &= P\Big(\bigcup_{j\in\Delta} \{\text{aceptar } H_j\}\Big) \\ &\leq \sum_{j\in\Delta} P(\text{aceptar } H_j) = \sum_{j\in\Delta} p_j \end{aligned} \tag{4}$$

(donde hemos usado la propiedad de sub-aditividad de la probabilidad para obtener la desigualdad), y aunque cada p_j con $j \in \Delta$ sea menor que α, no podemos asegurar que su suma lo sea. La solución propuesta por **Bonferroni** es muy sencilla: si cada uno de los m tests lo hacemos con nivel de

[1] El p-valor del test para contrastar la hipótesis alternativa H es, por definición, la probabilidad de aceptar H siendo falsa, esto es, p-valor $= P(\text{aceptar } H \,/\, H \text{ es falsa})$. Entonces, si el p-valor es menor que el nivel de significación que deseamos para el test (una cota superior, que denotamos por α, para la probabilidad de cometer el error de aceptar H siendo falsa, que se conoce como *error de tipo I*), es decir, si p-valor $< \alpha$, entonces se acepta la hipótesis H, ya que el error que se tiene de equivocarse será inferior al nivel de significación.

significación α/m, es decir, si para cada $i = 1, \ldots, m$ se acepta la hipótesis H_i si $p_i < \alpha/m$, entonces sí que podemos asegurar que la suma de los δ p-valores en (4) será $< \delta\,\alpha/m \leq \alpha$, ya que $\delta \leq m$ y, por tanto, la probabilidad del *family-wise error* seá $< \alpha$, que es lo que se pretendía.

En vez de modificar los niveles de significación con los que hacemos los m tests, podemos mantener siempre el mismo nivel de significación α a cambio de corregir (o *ajustar*) los p-valores: como $p_i < \alpha/m$ es equivalente a $m\,p_i < \alpha$, se definen los p-valores ajustados por el método de Bonferroni como $\widetilde{p}_i = m\,p_i$, y con estos p-valores ajustados las m comparaciones múltiples se realizan de la manera habitual: se acepta H_i si y sólo si $\widetilde{p}_i < \alpha$. Como los p-valores (sean o no ajustados) han de estar en el intervalo [0, 1], asumimos que el mayor de todos multiplicado por m es ≤ 1, lo que no es una restricción muy fuerte, es decir, asumimos que

$$p_m \leq \frac{1}{m}\,,$$

y así nos aseguramos de ello. Naturalmente, con este ajuste en los p-valores resulta más difícil cometer el *error de tipo I* (aceptar la hipótesis H_i siendo falsa), pues el rango de valores para p_i que permiten aceptar la hipótesis H_i es menor (es $[0, \alpha/m)$ en vez de ser $[0, \alpha)$) y, por el efecto balanza, será más probable cometer el *error de tipo II* (no aceptar H_i siendo cierta). Aunque este error se considera menos importante, tampoco interesa que su probabilidad sea muy alta y debe controlarse en la medida de lo posible.

Una mejor alternativa para solucionar el problema de las comparaciones múltiples es el método de **Holm-Bonferroni**, introducido por Holm en [8], que consigue que la probabilidad del *family-wise error* sea menor que el nivel de significación α pero aumenta menos que el método de Bonferroni el *error de tipo II* de los m tests. Esta variante del método original de Bonferroni se describe en el Algoritmo 1.

La idea de este algoritmo es: comparamos secuencialmente

$$p_1,\, p_2,\, p_3,\, \ldots,\, p_m \quad \text{con} \quad \frac{\alpha}{m},\, \frac{\alpha}{m-1},\, \frac{\alpha}{m-2}, \ldots,\, \frac{\alpha}{m-(m-1)} = \alpha$$

respectivamente, de manera que si $p_1 < \frac{\alpha}{m}$, se acepta H_1, y en caso contrario, se rechazan $H_1, H_2, \ldots, H_m$. Si se acepta H_1, se continúa con el siguiente paso: si $p_2 < \frac{\alpha}{m-1}$, se acepta H_2, y en caso contrario, se rechazan $H_2, H_3, \ldots, H_m$. Si se acepta H_2, se continúa con el siguiente paso: si $p_3 < \frac{\alpha}{m-2}$, se acepta H_3, y en caso contrario, se rechazan $H_3, \ldots, H_m$. Y así sucesivamente, de manera que si denotamos por m_0 el primer índice para el que no se acepta la correspondiente hipótesis, es decir, $m_0 = \text{mín}\{\, j = 1, \ldots, m \,:\, p_j \geq \frac{\alpha}{m-j+1}\,\}$, entonces $H_1, \ldots, H_{m_0-1}$ son aceptadas, y $H_{m_0}, \ldots, H_m$ son rechazadas. Con la notación que usamos en el algoritmo,

$$H_{acc} = \{H_1,\, \ldots,\, H_{m_0-1}\}$$

(se sobreentiente que si $m_0 = 1$, entonces $H_{acc} = \emptyset$).

Algoritmo 1 `Método de Holm-Bonferroni`

Input número de comparaciones múltiples $m \geq 2$, nivel de significación α,

p-valores $p_1 \leq p_2 \leq \cdots \leq p_m$ de las hipótesis $H_1, \ldots, H_m$, respectivamente

Output subconjunto de las hipótesis alternativas $H_1, \ldots, H_m$ aceptadas, H_{acc}

inicializamos $H_{acc} = \emptyset$
if $p_1 \geq \frac{\alpha}{m}$ **then**
  stop
else
  $H_{acc} = \{H_1\}$
  for j in $2 : m$ **do**
    if $p_j \geq \frac{\alpha}{m-j+1}$ **then**
      $H_{acc} = \{H_1, \ldots, H_{j-1}\}$
      stop
    else
      $H_{acc} = H_{acc} \cup \{H_j\}$
return H_{acc}

Proposición 2 *La probabilidad del family-wise error con el método de Holm-Bonferroni descrito en el Algoritmo 1 es* $< \alpha$.

Demostración:

La probabilidad del *family-wise error* es, como vimos en (4),

$$\begin{aligned} P(\text{aceptar alguna } H_j \text{ con } j \in \Delta) &= P\Big(\bigcup_{j \in \Delta} \{\text{aceptar } H_j\} \Big) \\ &\leq \sum_{j \in \Delta} P(\text{aceptar } H_j)\,, \end{aligned} \tag{5}$$

con la notación Δ introducida en (3), y vamos a demostrar que es $< \alpha$. En efecto, por construcción del método de Holm-Bonferroni, tal y como lo hemos descrito en el Algoritmo 1, que se acepte una hipótesis H_j con $j \in \Delta$ implica que se acepte H_{j_0}, siendo $j_0 = \text{mín } \Delta$, y esto sucede si

$$p_{j_0} < \frac{\alpha}{m - j_0 + 1}\,,$$

lo que quiere decir que para todo $j \in \Delta$,

$$P(\text{aceptar } H_j) \leq P(\text{aceptar } H_{j_0}) = p_{j_0} < \frac{\alpha}{m - j_0 + 1} \tag{6}$$

y por (5) y (6) tenemos que

$$\begin{aligned} P(\text{aceptar alguna } H_j \text{ con } j \in \Delta) &\leq \sum_{j \in \Delta} P(\text{aceptar } H_j) \\ &< \delta \, \frac{\alpha}{m - j_0 + 1} \, . \end{aligned} \tag{7}$$

Si $j_0 = 1$, la cota superior para la probabilidad del *family-wise error* en (7), $\delta \, \frac{\alpha}{m-j_0+1}$, es igual a $\delta \, \frac{\alpha}{m} \leq \alpha$ (ya que $\delta \leq m$), luego se acaba la demostración.

Si $j_0 > 1$, también por construcción del método de Holm-Bonferroni, que se acepte H_{j_0} con $j_0 \in \Delta$ implica que se acepten las hipótesis $H_1, \ldots, H_{j_0-1}$, pero como $\{H_1, \ldots, H_{j_0-1}\} \subset \Delta^c$, siendo $\Delta^c = \{1, \ldots, m\} \backslash \Delta$ el complementario de Δ en el conjunto $\{1, \ldots, m\}$, es decir, los índices de las hipótesis que son ciertas (dicho de otro modo, $H_1, \ldots, H_{j_0-1}$ son hipótesis ciertas) y se tiene que $j_0 - 1 \leq m - \delta$, ya que $m - \delta$ es el número total de hipótesis ciertas. Como consecuencia, $m - j_0 + 1 \geq \delta$ y la cota superior para la probabilidad del *family-wise error* en (7) será

$$\delta \, \frac{\alpha}{m - j_0 + 1} \leq \delta \, \frac{\alpha}{\delta} = \alpha \, ,$$

que es lo que queríamos demostrar. □

Al igual que sucede con el método de Bonferroni, para aplicar el método de Holm-Bonferroni, en vez de usar el Algoritmo 1, que implica modificar los niveles de significación con los que hacemos los m tests de manera secuencial, podemos mantener el nivel de significación α en todos ellos a cambio de ajustar los p-valores, que denotamos por $\widetilde{\widetilde{p}}_i$, $i = 1, \ldots, m$, así:

$$\widetilde{\widetilde{p}}_i = \text{máx}\{(m - j + 1) \, p_j, \, j \leq i\}$$

y las m comparaciones múltiples se realizan de la manera habitual: se acepta H_i si y sólo si $\widetilde{\widetilde{p}}_i < \alpha$. Aunque esta fórmula puede parecer complicada, en realidad no lo es. La idea es simple: para el p-valor menor, p_1, como compararlo con α/m es equivalente a multiplicarlo por m y comparar con α, tenemos que $\widetilde{\widetilde{p}}_1 = m \, p_1$. A partir del segundo p-valor más pequeño, el procedimiento tiene dos etapas. Por ejemplo, para ajustar p_2:

1) multiplicamos p_2 por $m - 1$. Si este producto es mayor que $\widetilde{\widetilde{p}}_1$, que es el p-valor ajustado anterior, se toma $\widetilde{\widetilde{p}}_2 = (m - 1) \, p_2$,

2) en caso contrario, el p-valor ajustado se define como el anterior, $\widetilde{\widetilde{p}}_2 = \widetilde{\widetilde{p}}_1$. De esta manera se consigue que si se acepta H_2, necesariamente se acepten las hipótesis anteriores, en este caso, H_1.

Y así sucesivamente. La Tabla 16 nos muestra un ejemplo de ajuste de p-valores por los dos métodos, Bonferroni y Holm-Bonferroni.

p-valores p_i	$\widetilde{p}_i$ (Bonferroni)	rango $m-i+1$	$(m-i+1)\,p_i$	$\widetilde{\widetilde{p}}_i$ (Holm-Bonferroni)
0.001	$0.001 \times 5 = 0.005$	5	$0.001 \times 5 = 0.005$	0.005
0.005	$0.005 \times 5 = 0.025$	4	$0.005 \times 4 = 0.020$	0.020
0.015	$0.015 \times 5 = 0.075$	3	$0.015 \times 3 = 0.045$	0.045
0.020	$0.020 \times 5 = 0.100$	2	$0.020 \times 2 = 0.040$	0.045
0.100	$0.100 \times 5 = 0.500$	1	$0.100 \times 1 = 0.100$	0.100

Tabla 16: Ejemplo con $m = 5$ de obtención de los p-valores ajustados usando los métodos de Bonferroni y de Holm-Bonferroni.

Ambos métodos tienen una probabilidad del *family-wise error* $< \alpha$, pero con nivel de significación $\alpha = 0.05$, por ejemplo, mientras que con el método de Bonferroni sólo aceptaríamos las hipótesis alternativas H_1 y H_2, con el método de Holm-Bonferroni aceptaríamos H_1, H_2, H_3 y H_4. De esta manera, vemos que la probabilidad de cometer el *error de tipo II* (rechazar una hipótesis alternativa H siendo cierta) es menor con el método de Holm-Bonferroni, ya que sus p-valores ajustados son menores que con el método de Bonferroni, es decir, que

$$\text{para todo } i = 1, \ldots, m\,, \quad p_i \leq \widetilde{\widetilde{p}}_i \leq \widetilde{p}_i \leq 1\,.$$

Apéndice C: Tablas de probabilidad condicionada para la variable *Resultado*

Cada una de las tablas de probabilidad condicionada (CPT, por *Conditional Probability Table*) de esta sección se ha obtenido estimando con el modelo EWA aprendido a partir de toda la base de datos, la probabilidad *a posteriori* de la variable *Resultado* a cada uno de los posibles valores de las características F_1 a F_{21}, suponiendo que el resto de características no han sido observadas, excepto en el caso de las características de la categoría "Principal causa de ingreso", para las que cuando una de ellas es presente, las otras necesariamente han de estar ausentes ya que son mútuamente excluyentes[2]

	F_1: Sexo		F_2: Edad					
	Hombre	Mujer	< 45	45–54	55–64	65–74	75–84	> 84
vivo	**0.86575**	**0.83143**	**0.90517**	**0.89967**	**0.89313**	**0.84506**	**0.80894**	**0.82340**
muerto	0.13425	0.16857	0.09483	0.10033	0.10687	0.15494	0.19106	0.17660

Tabla 17: CPT de la variable *Resultado* conditionada a F_1 y a F_2.

[2] Aunque es posible que un paciente presente más de una de las características de esta categoría, de F_4 a F_{16} en la práctica, en principio sólo es recogida la que el especialista considera más importante, y por ello se puede asumir que son mútuamente excluyentes.

	F_3: índice de Charlson					F_4: ACS	
	0	1	2	3	> 3	no	sí
vivo	**0.90627**	**0.86975**	**0.84657**	**0.81614**	**0.76523**	**0.82243**	**0.99610**
muerto	0.09373	0.13025	0.15343	0.18386	0.23477	0.17757	0.00390

Tabla 18: CPT de la variable *Resultado* conditionada a F_3 y a F_4.

	F_5: RF		F_6: Shock		F_7: Coma		F_8: Renal F	
	no	sí	no	sí	no	sí	no	sí
vivo	**0.87220**	**0.83919**	**0.88342**	**0.79644**	**0.86008**	**0.83497**	**0.85822**	**0.84957**
muerto	0.12780	0.16081	0.11658	0.20356	0.13992	0.16503	0.14178	0.15043

Tabla 19: CPT de la variable *Resultado* condicionada a F_5, a F_6, a F_7 y a F_8.

	F_9: Hepatic F		F_{10}: CRA		F_{11}: ES		F_{12}: Arrhythmia	
	no	sí	no	sí	no	sí	no	sí
vivo	**0.85302**	**0.98666**	**0.87396**	0.37702	**0.84377**	**0.99580**	**0.85084**	**0.94933**
muerto	0.14698	0.01334	0.12604	**0.62298**	0.15623	0.00420	0.14916	0.05067

Tabla 20: CPT de la variable *Resultado* condicionada a F_9, a F_{10}, a F_{11} y a F_{12}.

	F_{13}: CT		F_{14}: OT		F_{15}: Intoxication		F_{16}: Other syndromes	
	no	sí	no	sí	no	sí	no	sí
vivo	**0.85369**	**0.69640**	**0.85177**	**0.97827**	**0.85271**	**0.92872**	**0.84672**	**0.96777**
muerto	0.14631	0.30360	0.14823	0.02173	0.14729	0.07128	0.15328	0.03223

Tabla 21: CPT de la variable *Resultado* condicionada a F_{13}, a F_{14}, a F_{15} y a F_{16}.

	F_{17}: Origen					
	Planta	Quirófano	Emergencias	Emergencias extra-hospitalarias	Otro hospital	Desconocido
vivo	**0.77059**	**0.89578**	**0.88156**	**0.53767**	**0.87632**	**0.71747**
muerto	0.22941	0.10422	0.11844	0.46233	0.12368	0.28253

Tabla 22: CPT de la variable *Resultado* condicionada a F_{17}.

	F_{18}: Síndrome genérico					
	Cirugía electiva	Cirugía urgente	Coronario	Médico	Trauma	Desconocido
vivo	**0.94701**	**0.85312**	**0.97963**	**0.80451**	**0.93671**	**0.93971**
muerto	0.05299	0.14688	0.02037	0.19549	0.06329	0.06029

Tabla 23: CPT de la variable *Resultado* condicionada a F_{18}.

	F_{19}: Septicemia	
	no	sí
vivo	**0.89398**	**0.78065**
muerto	0.10602	0.21935

Tabla 24: CPT de la variable *Resultado* condicionada a F_{19}.

	F_{20}: Carga de trabajo UCI					
	Seguimiento méd.	Inestable con coma/con.	Inestable sin coma/con.	Seguimiento post-oper.	Inestable post-oper.	Desc.
vivo	**0.95564**	**0.55982**	**0.86603**	**0.99195**	**0.98728**	**0.85176**
muerto	0.04436	0.44018	0.13397	0.00805	0.01272	0.14824

Tabla 25: CPT de la variable *Resultado* condicionada a F_{20}.

	F_{21}: APACHE II								
	< 5	5–9	10–14	15–19	20–24	25–29	30–34	> 34	Desconocido
vivo	**0.99579**	**0.96505**	**0.92690**	**0.86155**	**0.71854**	**0.59862**	**0.58245**	0.36033	**0.76492**
muerto	0.00421	0.03495	0.07310	0.13845	0.28146	0.40138	0.41755	**0.63967**	0.23508

Tabla 26: CPT de la variable *Resultado* condicionada a F_{21}.

Referencias

[1] Bi, Y., Guan, J., Bell, D. The combination of multiple classifiers using an evidential reasoning approach. Artificial Intelligence, 172(15), pp. 1731–1751 (2008).

[2] Delgado, R. A semi-hard voting combiner scheme to ensemble multi-class probabilistic classifiers. Appl Intell (2021). https://doi.org/10.1007/s10489-021-02447-7

[3] Delgado, R. Xarxes Bayesianes: una metodologia per avaluar riscos. Nou Biaix-Revista de la FEEMCAT i la SCM, 45, pp. 4–30 (Diciembre 2019).

[4] Delgado, R. Derecho y Probabilidad: Falacias, Fórmula de Bayes y Redes Bayesianas. MATerials MATemàtics, trabajo no. 6, 31 pp. ISSN: 1887-109 (2013). http://mat.uab.cat/web/matmat/v2013n06/

[5] Delgado, R., Núñez-González, J.D.,Yébenes, J.C., Lavado, A. Survival in the Intensive Care Unit: A prognosis model based on Bayesian classifiers. Artificial Intelligence in Medicine, 115 (2021) 102054, ISSN 0933-3657. https://doi.org/10.1016/j.artmed.2021.102054.

[6] Detsky, M.E., Harhay, M.O., Bayard, D.F., Delman, A.M., Buehler, A.E., Kent, S.A., Ciuffetelli, I.V., Cooney, E., Gabler, N.B., Ratcliffe, S.J., Mikkelsen, M.E., Halpern, S.D. Six-Month Morbidity and Mortality among Intensive Care Unit Patients Receiving Life-Sustaining Therapy. A Prospective Cohort Study. Ann Am Thorac Soc, 14(10), pp. 1562–1570 (2017).

[7] Granholm, A., Miller, M.H., Krag, M., Perner, A., Hjortrup, P.B. Predictive Performance of the Simplified Acute Physiology Score (SAPS) II and the Initial Sequential Organ Failure Assessment (SOFA) Score in Acutely Ill Intensive Care Patients: Post-Hoc Analyses of the SUP-ICU Inception Cohort Study. PLoS One, 11(12): e0168948 (2016). https://doi.org/10.1371/journal.pone.0168948

[8] Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics, 6(2), pp. 65–70 (1979).

[9] Kerlin, M.P., Cooke, C.R. Understanding Costs When Seeking Value in Critical Care. Ann Am Thorac Soc, 12(12), pp. 1743–1744 (2015).

[10] Kittler, J., Hatef, M., Duin, R.P.W., Matas, J. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3), pp. 226–239 (1998).

[11] Knaus, W.A., Draper, E.A., Wagner, D.P., Zimmerman, J.E. APACHE II: a severity of disease classification system. Crit Care Med, 13(10), pp. 818–829 (1985).

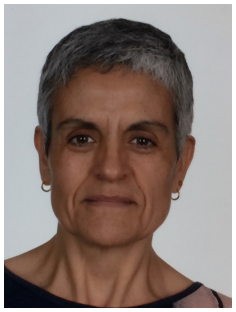
[12] Kuncheva, L.I. Combining Pattern Classifiers. John Wiley & Sons, Inc. (2004).

[13] Li, Z., Cheng, B., Wang, J., Xie, G., Yu, X., Huang, M., Xu, Z., Lu, Z., Sun, H., Zhang, J., Wang, Z., Wu, H., Liu, X., Chu, L., Zhao, J., Fang, X. A multifactor model for predicting mortality in critically ill patients: A multicenter prospective cohort study. J Crit Care, 42, pp. 18–24 (2017).

[14] Lone, N.I., Gillies, M.A., Haddow, C., Dobbie, R., Rowan, K.M., Wild, S.H., Murray, G.D., Walsh, T.S. Five-Year Mortality and Hospital Costs Associated with Surviving Intensive Care. Am J Respir Crit Care Med, 194(2), pp. 198–208 (2016).

[15] Niewiński, G., Starczewska, M., Kański, A. Prognostic scoring systems for mortality in intensive care units. The APACHE model. Anaesthesiol Intensive Ther, 46(1), pp. 46–49 (2014).

[16] Xu, L., Krzyzak, A., Suen, C.Y. Methods of combining multiple classifiers and their application to handwriting recognition. IEEE Transactions on Systems, Man, and Cybernetics, 22, pp. 418–435 (1992).



Departament de Matemàtiques
Universitat Autònoma de Barcelona
delgado@mat.uab.cat

Publicat el 2 de desembre de 2021