



ORIGINAL

¿Podrían ayudarnos los algoritmos de *machine learning* en la predicción de hemorragia masiva a nivel prehospitalario?



Marcos Valiente Fernández^{a,*}, Carlos García Fuentes^a,
Francisco de Paula Delgado Moya^a, Adrián Marcos Morales^a,
Hugo Fernández Hervás^a, Jesús Abelardo Barea Mendoza^a,
Carolina Mudarra Reche^a, Susana Bermejo Aznárez^a, Reyes Muñoz Calahorra^a,
Laura López García^a, Fernando Monforte Escobar^b y Mario Chico Fernández^a

^a UCI de Trauma y Emergencias, Hospital Universitario 12 de Octubre, Madrid, España

^b Servicio de Asistencia Municipal de Urgencia y Rescate (SAMUR)-Protección Civil, Madrid, España

Recibido el 15 de abril de 2023; aceptado el 21 de junio de 2023

Disponible en Internet el 18 de agosto de 2023

PALABRAS CLAVE

Trauma;
Prehospitalaria;
Scores clínicos;
Machine learning;
Hemorragia masiva

Resumen

Objetivo: Comparación de la capacidad predictiva de diferentes algoritmos de *machine learning* (AML) respecto a escalas tradicionales de predicción de hemorragia masiva en pacientes con enfermedad traumática grave (ETG).

Diseño: Sobre una base de datos de una cohorte retrospectiva con variables clínicas prehospitalarias y de resultado de hemorragia masiva se realizó un tratamiento de la base de datos para poder aplicar los AML, obteniéndose un conjunto total de 473 pacientes (80% entrenamiento, 20% validación). Para la modelización se realizó imputación proporcional y validación cruzada. El poder predictivo se evaluó con la métrica ROC y la importancia de las variables mediante los valores Shapley.

Ámbito: Atención extrahospitalaria del paciente con ETG.

Pacientes: Pacientes con ETG atendidos en el medio extrahospitalario por un servicio médico extrahospitalario desde enero de 2010 hasta diciembre de 2015 y trasladados a un centro de trauma en Madrid.

Intervenciones: Ninguna.

Variables de interés principales: Obtención y comparación de la métrica ROC de 4 AML: *random forest*, *support vector machine*, *gradient boosting machine* y *neural network* con los resultados obtenidos con escalas tradicionales de predicción.

* Autor para correspondencia.

Correo electrónico: mvalientefernandez@gmail.com (M. Valiente Fernández).

Resultados: Los diferentes AML alcanzaron valores ROC superiores al 0,85, teniendo medianas cercanas a 0,98. No encontramos diferencias significativas entre los AML. Cada AML ofrece un conjunto de variables diferentes, pero con predominancia de las variables hemodinámicas, de reanimación y de deterioro neurológico.

Conclusiones: Los AML podrían superar a las escalas tradicionales de predicción en la predicción de hemorragia masiva.

© 2023 Elsevier España, S.L.U. y SEMICYUC. Todos los derechos reservados.

KEYWORDS

Trauma;
Prehospital;
Clinical scores;
Machine learning;
Massive hemorrhage

Could machine learning algorithms help us in predicting massive hemorrhage at the prehospital level?

Abstract

Objective: Comparison of the predictive ability of various machine learning algorithms (MLA) versus traditional prediction scales for massive hemorrhage in patients with severe traumatic injury (ETG).

Design: On a database of a retrospective cohort with prehospital clinical variables and massive hemorrhage outcome, a treatment of the database was performed to be able to apply the different MLA, obtaining a total set of 473 patients (80% training and 20% validation). For modeling, proportional imputation and cross validation were performed. The predictive power was evaluated with the ROC metric and the importance of the variables using the Shapley values.

Setting: Out-of-hospital care of patients with ETG.

Participants: Patients with ETG treated out-of-hospital by a prehospital medical service from January 2010 to December 2015 and transferred to a trauma center in Madrid.

Interventions: None.

Main variables of interest: Obtaining and comparing the ROC curve metric of 4 MLAs: random forest, support vector machine, gradient boosting machine and neural network with the results obtained with traditional prediction scales.

Results: The different MLA reached ROC values higher than 0.85, having medians close to 0.98. We found no significant differences between MLAs. Each MLA offers a different set of more important variables with a predominance of hemodynamic, resuscitation variables and neurological impairment.

Conclusions: MLA may be helpful in patients with massive hemorrhage by outperforming traditional prediction scales.

© 2023 Elsevier España, S.L.U. y SEMICYUC. All rights reserved.

Introducción

La enfermedad traumática grave (ETG) es la principal causa de muerte en menores de 40 años (en 2010, 5,1 millones de muertes, una de cada 10 muertes fue por enfermedad traumática)¹. Dentro del ámbito de la ETG la hemorragia masiva (HM) es la principal causa de muerte potencialmente prevenible². La detección precoz de estos pacientes y el uso de protocolos de transfusión masiva (PTM), junto con la reanimación con control de daños, han permitido que la mortalidad se haya reducido a un 30%³. Para detectar pacientes con HM se han desarrollado diferentes escalas predictivas y se han realizado estudios para identificar las escalas tradicionales predictivas (ETP) de hemorragia masiva más útiles en nuestro medio⁴⁻⁶. Estas ETP usan la estadística clásica (regresión lineal o logística) y están sujetas a un conjunto de restricciones matemáticas. Actualmente la estadística computacional se está haciendo más relevante por su mayor capacidad predictiva gracias a los algoritmos de *machine learning* (AML) y su diferente enfoque de trabajo: trabajan

sobre los datos buscando patrones lineales y no lineales y con menos restricciones^{7,8}, ya que no buscan adaptar los datos a condiciones matemáticas (homocedasticidad, por ejemplo), sino que buscan optimizar la capacidad predictiva con riesgo de sobreajuste y menor grado de explicabilidad. La necesidad de detectar mejor y más precozmente a los pacientes con HM nos obliga a explorar alternativas a las ETP⁹.

El objetivo es explorar la capacidad predictiva que ofrecen los AML frente a las ETP en la detección de aparición de futuras HM en adultos que han sufrido una ETG. En caso de observar una gran capacidad predictiva indagaríamos en la importancia de las variables de los diferentes AML.

Métodos

Base de datos

Debido a que conocemos qué pacientes tuvieron el evento de interés (presentar HM) y que la intención del estudio es

determinar la capacidad predictiva de HM, aplicaremos AML tipo «aprendizaje supervisado».

Se trata de un estudio de cohortes retrospectivo. La base de datos empleada es propia, no financiada, confidencial, que recoge sistemática y diariamente diferentes aspectos de la enfermedad traumática. Incluye a los pacientes mayores de 15 años con trauma grave, atendidos desde enero de 2010 hasta diciembre de 2015 por SAMUR-Protección Civil (PC), el servicio de emergencias sanitarias de la ciudad de Madrid, y trasladados posteriormente al Hospital 12 de Octubre, centro hospitalario de alta complejidad en dicha ciudad. El origen de los datos es mixto, siendo los de la asistencia prehospitalaria extraídos de una base de datos que SAMUR-PC elabora de manera prospectiva y con otra base de datos prospectiva hospitalaria del Hospital 12 de Octubre. Se trata de la misma muestra poblacional usada en el estudio *Predicción de hemorragia masiva a nivel extrahospitalario: validación de seis escalas*⁵. El estudio⁵ fue aprobado por el Comité de Ética.

Se incluyeron consecutivamente todos aquellos pacientes con trauma grave definido por criterios fisiológicos y anatómicos. Se excluyeron los pacientes que a su ingreso en el hospital: 1) se encontraban en situación de parada cardiorrespiratoria o situación *pre mortem* y se desestimaron las maniobras de reanimación; 2) rechazo del paciente o familiares a recibir productos sanguíneos o derivados. La atención inicial extrahospitalaria se realiza por un equipo de atención especializado, constituido por 2 médicos, 2 enfermeras y 2 técnicos según los esquemas del *Advanced Trauma Life Support* y un procedimiento de asistencia conjunto denominado Código Trauma¹⁰. Los datos se recogen de forma prospectiva en una base de datos hospitalaria propia específica de trauma. Se decidió estudiar las siguientes variables: variables demográficas (edad y sexo), variables clínicas (mecanismo lesional, sospecha de fractura de fémur o de pelvis inestable), variables fisiológicas (la primera frecuencia cardiaca, la tensión arterial sistólica [TAS] y la tensión arterial diastólica), variables analíticas (hemoglobina, exceso de bases y ácido láctico de sangre venosa [analyzer de gases en sangre epoc[®], de Siemens Healthineers]), variables de tratamiento extrahospitalario (cantidad de fluidos prehospitalarios administrados), variables de imagen (se realizó una ecografía tipo *Focused Abdominal Sonography for Trauma*, para lo que se utilizó un ecógrafo portátil tipo Sonosite 180PLUS[®] de FujiFilm [variables recogidas del informe estructurado del SAMUR-PC], variables de seguimiento hospitalario en UCI (lesiones anatómicas, gravedad [*Injury Severity Score*], necesidad de arteriografía/cirugía para control de hemorragia [variables recogidas de la base de datos de nuestra unidad]) y hemocomponentes trasfundidos. La variable de resultado principal, que va a intentar ser predicha, es la HM.

Criterio de hemorragia masiva

El objetivo del estudio es la predicción de HM, por lo que se realizó una consulta de nuestro registro de transfusión, de donde se recogieron el número de concentrados de hemáties transfundidos a cada paciente. En este estudio se ha definido el criterio de HM con la misma definición que el estudio comparativo⁵: administración de ≥ 10 concentrados

de hemáties en las primeras 24 h de ingreso tras el trauma (≥ 2.500 cc). En nuestra base de datos, tras el preprocesado hemos obtenido un total de casos de 7,9%, algo menor que el estudio comparativo (9,2%), lo que, en principio, dificultaría aún más la predicción por su infrecuencia.

Limpieza, reorganización y preprocesamiento

Ha de quedar claro que en este trabajo vamos a comparar los resultados obtenidos previamente por las ETP en una cohorte de pacientes (*Predicción de hemorragia masiva a nivel extrahospitalario: validación de seis escalas*⁵) con los resultados obtenidos por los AML en un subgrupo (conjunto de validación) que proviene de esa misma cohorte.

En el estudio comparativo⁵ se calcularon con las variables fisiológicas, anatómicas, analíticas y de imagen del medio extrahospitalario las siguientes escalas predictivas de HM: 1) *Trauma Associated Severe Haemorrhage Score*^{11,12}; 2) *Assessment of Blood Consumption Score*¹³; 3) *Emergency Transfusion Score*¹⁴; 4) *índice de shock (IS)*^{15,16}; 5) *Prince of Wales Hospital/Rainer Score*⁴, y 6) *Larson Score*¹⁷. Las ETP con las que vamos a comparar los AML¹⁸ son:

El preprocesado de datos consistió en las siguientes acciones¹⁹:

1. Retirada de variables por pérdidas significativas de datos (numerador = datos disponibles, denominador = datos totales): exceso de bases (351/550), lactato (169/550), hemoglobina (491/550), hematocrito (141/550); además, se retiraron valores ecográficos al no haber sido realizados sistemáticamente. Debido a las dificultades, a veces existentes, de codificar el mecanismo lesional se prescindió también de este. Es destacable que estamos eliminando variables muy usadas en la toma de decisiones del día a día.
2. Retirada de variables que no pertenecen al campo prehospitalario: *Injury Severity Score*, Estancia, Evolución.
3. En variables con pérdidas menores al 5% se imputaron datos mediante el paquete MICE, que sigue la metodología *Multiple Imputations by Chained Equations*.
4. En resumen, hemos retirado variables con muchas pérdidas y que no son del contexto de estudio. Tras la selección de variables hemos completado las observaciones de variables con pocas pérdidas. Sin embargo, observamos que la variable de interés (HM) es muy desbalanceada (< 10% de los casos). Estos problemas de desequilibrio y de tamaño muestral relativamente escaso generan problemas de aprendizaje algorítmico, por lo que se decidió, para facilitar el aprendizaje, que se utilizaran técnicas de síntesis/generación de casos mediante el paquete *Synthetic Minority Over-sampling Technique*, –SMOTE– manteniendo la proporcionalidad de casos.

Selección de variables

Tras el preprocesamiento las variables recopiladas fueron: sexo, edad, puntuación en la *Glasgow Coma Scale*, frecuencia cardiaca, tensión arterial diastólica inicial, TAS inicial, volumen de fluidos administrados prehospitalariamente, sospecha de lesión pélvica y/o de fémur, presencia de trauma penetrante. Estas variables permitieron la crea-

ción de «variables derivadas», que permiten condensar más información en una variable, tales como: tensión arterial media inicial e IS (prehospitalario) y agrupación del nivel neurológico en *Glasgow Coma Scale* bajo, intermedio o alto.

Desarrollo de modelos: creación de *datasets* de entrenamiento y validación

Los pasos que se han seguido para la aplicación de los AML han sido los siguientes:

- 1) Preprocesado del conjunto de datos (selección de variables, imputación y síntesis) y derivación de variables (IS) de la base de datos utilizada en el estudio previo para la aplicación de AML. Se comprueba gráficamente que tras la síntesis de casos se mantiene una distribución de la variable de interés en el espacio muestral similar en los datos originales y sintéticos (ver [material suplementario](#)).
- 2) Partición aleatoria del conjunto de datos en conjunto de entrenamiento (80% de los casos) y conjunto de validación (20% restante).
- 3) Optimización algorítmica: detección del mejor modelo (precisión y consistencia) mediante selección de hiperparámetros, validación cruzada con remuestreo y repetición.

Selección del modelo y ajuste hiperparamétrico

Los AML utilizados fueron: *random forest* (RF), *support vector machine* (SVM), *neural network* (NN) y *gradient boosting machine* (GBM) ([tabla 1](#)).

Los algoritmos buscan la mejor solución (predicción) a un problema dado un conjunto de datos²⁰. La búsqueda de esta solución requiere de una metodología, siendo esta diferente en cada algoritmo. El «método» que sigue cada algoritmo se define con los hiperparámetros. Por ejemplo, en el algoritmo de RF uno de los hiperparámetros sería el número de árboles con el que se toma una decisión. Como sus valores óptimos cambian según el conjunto de datos, se realiza una exploración de diferentes valores de estos hiperparámetros. Es decir, aplicamos diferentes algoritmos que «exploran» el conjunto de datos de manera diferente, e hiperparámetros que permiten una toma de decisiones diferentes sobre esa exploración de datos²¹.

Evaluación del rendimiento de los modelos matemáticos

La métrica *area under the receiver operator characteristics* (AUROC) ha sido la métrica guía en la comparación; además, se comparan otras métricas derivadas de la matriz de confusión. Como ya se ha comentado, la gran diferencia es que los resultados de las ETP recogidas en el estudio previo⁵ se realizan frente a toda la cohorte, mientras que los AML los enfrentamos contra un subconjunto que cuenta con observaciones sintéticas (conjunto test/validación).

La obtención de resultados estadísticos se realizó mediante el uso de R 4.1.2 y RStudio 1.3.1093. Las librerías usadas quedan reflejadas en el [material suplementario](#).

El resumen de todo el proceso queda reflejado en la [figura 1](#).

Valoración de la importancia de las variables en cada algoritmo

Se exploró la importancia de las variables en cada AML ([fig. 2](#)).

Resultados

En relación con el objetivo principal, se obtuvo la métrica ROC de los diferentes AML frente al conjunto de entrenamiento: RF (AUC 0,993, intervalo de confianza [IC]: 0,992-0,994), SVM (AUC 0,99, IC: 0,995-0,997), GBM (AUC 0,98, IC: 0,97-0,982) y NN (AUC 0,99, IC: 0,98-0,991). Este valor promedio de la curva ROC es un estimado que se obtiene de realizar múltiples mediciones mediante validación cruzada (10 particiones, 100 repeticiones, por lo tanto 1.000 mediciones) en el conjunto de datos de entrenamiento ([fig. 3](#)). Además, se expresan las siguientes métricas: sensibilidad, especificidad, valores predictivos positivo y negativo y razones de verosimilitud positiva y negativa.

El estimado se obtiene gracias a la técnica de validación cruzada, que sirve para generar consistencia en la estimación (se aprecia en los valores de los diferentes IC). Estos resultados se obtienen con el conjunto de datos de entrenamiento, pero la verdadera validación se realiza con el conjunto de datos de test, objetivando muy buenos resultados en la matriz de confusión (con unos valores predictivos positivos mayores) y en la diferencia de acierto entre el conjunto de entrenamiento y el conjunto test ([fig. 4](#)). No encontramos diferencias significativas entre los diferentes AML de cara a la predicción.

Los AML nos permiten investigar la importancia de cada variable en la toma de decisiones ([fig. 2](#)). Como cada AML tiene reglas internas diferentes, arrojan distintos resultados.

Discusión

Los datos obtenidos muestran una alta capacidad predictiva de los diferentes AML con AUC más elevadas que las reflejados por las ETP en el estudio comparativo⁵. Los resultados obtenidos por las ETP, al estar enfrentados a pacientes reales, se podrían considerar más consistentes, y los del AML, al estar enfrentados a pacientes reales y sintéticos, se podrían considerar más inestables y requerirían de un estudio de validación. Debemos reconocer que no se establece una comparación estadística concreta para saber si realmente son mejores los AML que los ETP, sin embargo, observamos indicios de que así podría ser: IC de los diferentes AML elevados y distantes de los resultados de los AUC de las ETP. Los valores de sensibilidad y especificidad que corresponden al mejor punto de corte de la curva ROC de los AML, presentan valores predictivos positivos y negativos y razones de verosimilitud positiva mayores que los de las ETP ([figs. 3 y 4](#)).

Asimismo, cada AML nos ofrece un conjunto de variables más importantes, no siempre plenamente coincidentes.

Tabla 1 Explicación general de los algoritmos de *machine learning*

Modelo	Idea intuitiva	Hiperparámetros
<i>Support vector machine</i>	Algoritmo que aumenta la dimensionalidad de los datos y busca, mediante hiperplanos (elementos que permiten la separación de un conjunto de datos), la zona que mejor discrimina. Son algoritmos que se desarrollaron para clasificaciones binarias, las cuales son frecuentes en el ámbito clínico	Sigma: permite rebajar las condiciones impuestas por la fórmula, lo que da «holgura». Es por ello que valores más altos pueden producir sobreajuste Coste: penalización por «uso» del margen <i>soft</i> debido a mala clasificación
<i>Neural network</i>	Son algoritmos basados en las funciones biológicas neuronales. Se trata de un sistema de «neuronas densamente conectadas» que reciben información (valores de variables), las cuales transforman (sumatorios y funciones de activación) para ofrecer un resultado de salida (clase/resultado) Cuando el resultado no coincide con lo esperado tratan de corregir el resultado ajustando las transformaciones de adelante a atrás (<i>back propagation</i>)	Size: es el número de componentes (neuronas) que hay en la capa oculta Decay: es el parámetro que determina la penalización/regularización con intención de evitar el sobreajuste
<i>Random forest</i>	Generación de múltiples árboles de decisión aleatorios mediante técnicas de <i>bootstrapping (random)</i> , de tal manera que la predicción se realiza por la mayoría del resultado de los árboles (el ensamble [agregación de resultados] mediante <i>aggregating</i> es lo que forma el <i>forest</i>). Es, por lo tanto, una técnica de <i>ensemble</i> (agrupación) de árboles de decisión aleatorios La generación de múltiples árboles aleatorios diferentes y su agregación conjunto es lo que se conoce como <i>bagging (bootstrapping + aggregating)</i> . Esta combinación de múltiples modelos en uno nuevo se realiza con el objetivo de lograr un equilibrio entre sesgo y varianza, consiguiendo así mejores predicciones	Mtry: número de variables que pueden dividir en cada nodo Splitrule: normas que rigen la división de los nodos Min.node.size: valor que determina el tamaño más pequeño del nodo Importance: determinamos el criterio de cómo medimos el grado de «uniformidad» de cada nodo. Es decir, la mejoría tras cada división
<i>Gradient boosting machine</i>	La generalización de la técnica de <i>boosting</i> (ajuste de modelos basados en <i>weak learners</i> tras cada iteración), mediante el empleo de cualquier función de coste que sea diferenciable, permitió implementar el <i>gradient boosting</i> a multitud de problemas Al algoritmo de <i>gradient boosting</i> se le incorporaron propiedades de <i>bagging</i> implementando en cada iteración del algoritmo que el <i>weak learner</i> se ajuste empleando únicamente una fracción del set de entrenamiento, extraída de forma aleatoria (<i>stochastic</i>), desarrollando así el <i>stochastic gradient boosting</i> , que es el algoritmo que recoge carret en el método GBM. Este algoritmo consigue mejorar la capacidad predictiva y permite estimar el <i>out-of-bag-error</i> bajo ciertas circunstancias	Ntrees: número de árboles para realizar el ajuste. Normalmente utiliza muchos árboles ya que aunque puede sobreajustar, luego se busca el número mínimo de árboles que minimice la función de pérdida Interaction.depth: hace referencia a la profundidad del árbol. El valor <i>default</i> es 1, que significa que se utiliza como <i>weak learner</i> el <i>decision stump</i> del árbol de decisión Shrinkage: es la tasa de aprendizaje (<i>learning rate</i>). Cómo de rápido realiza el algoritmo el descenso de gradiente. A mayor número, menos sobreajuste pero mayor coste computacional n.minobsinnode: mínimo número de observaciones permitidas en los nodos terminales de los árboles

Generalmente incluyen criterios hemodinámicos (tensión arterial, frecuencia y su relación), de reanimación (fluidoterapia) y neurológicos (presencia de deterioro neurológico). Aquí exponemos como ejemplo los arrojados por la red neuronal (fig. 5), apreciándose que la variable más importante en la predicción es: IS > fluidos > frecuencia cardiaca > GCS > otras variables. Si hubiese un AML mejor que el resto para este problema puede que la importancia de las

variables nos ayudase a trasladar su explicabilidad al ámbito clínico.

Como hallazgos clínicos relevantes podríamos encontrar 2: 1) destaca que no aparece la TAS (como variable aislada) sistemáticamente como una de las variables más importantes, lo que puede deberse a que la TAS ayuda a detectar el shock pero son otras las variables que nos matizan del grado de shock y es por ello que quizá sea mucho más relevante

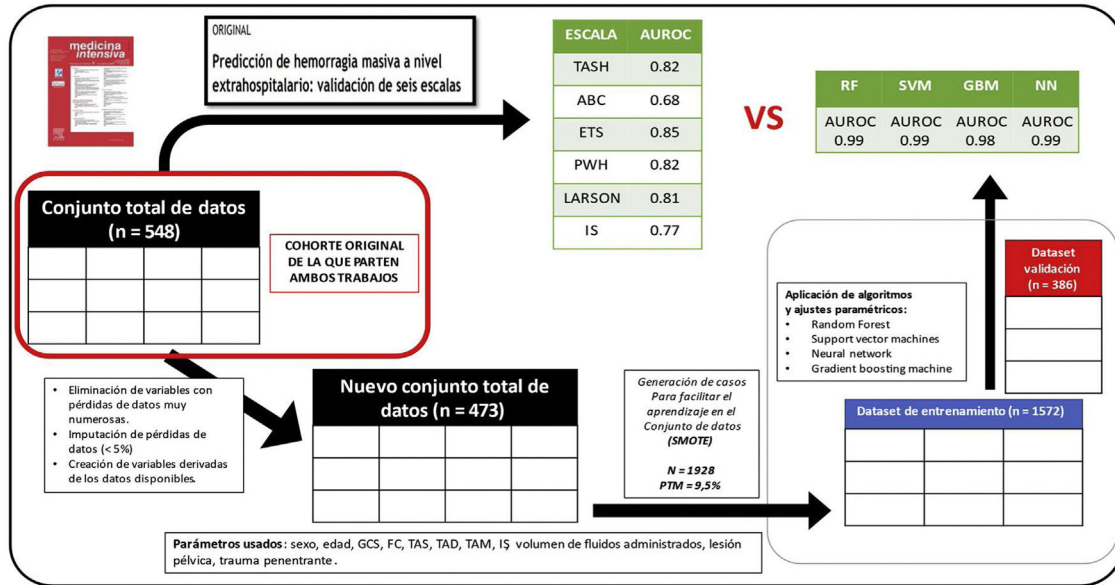


Figura 1 Visión general del preprocesamiento de datos y comparativa de resultados. Representación gráfica de las fases de preprocesamiento, modelización, obtención de resultados y comparación de resultados con estudio previo.

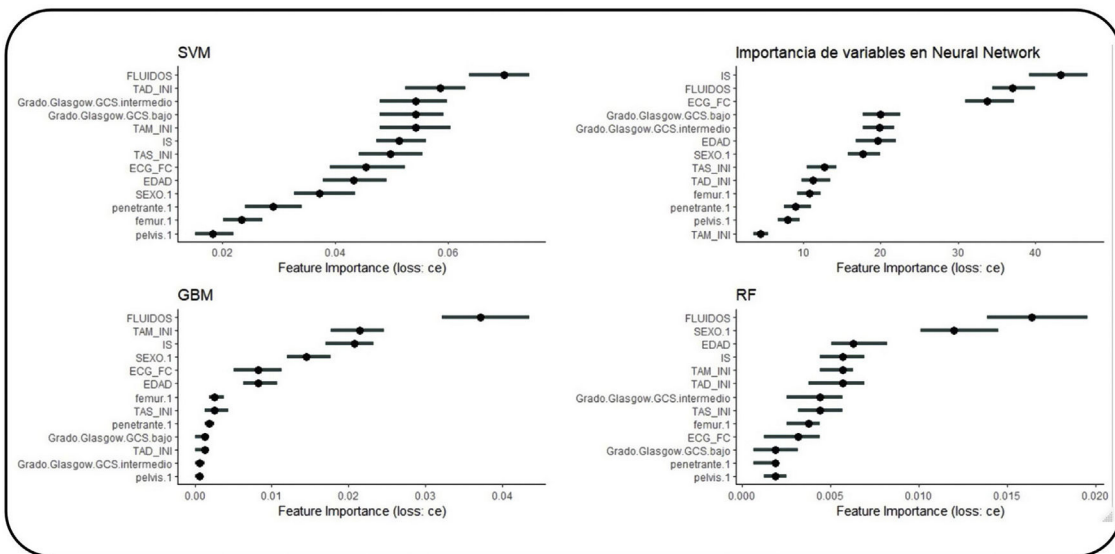


Figura 2 Importancia de las variables en cada algoritmo.

el IS que la TAS en sí misma, y 2) encontramos la presencia de deterioro neurológico como predictor de HM. Este mismo hecho sólo quedaba reflejado dentro de las ETP-HM en la escala *Prince of Wales Hospital/Rainer Score*. Presenta sentido clínico ya que podría reflejar estados de hipoperfusión cerebral graves secundarios a la hemorragia. Este hecho podría generar el debate sobre la monitorización de pacientes con shock hemorrágico y discreta lesión anatómica cerebral, ya que más que traducir lesión cerebral podría referir shock profundo.

Desde el punto de vista metodológico es resaltable en este artículo la alta capacidad predictiva de los AML utilizando únicamente variables «básicas» (sin utilizar mediciones bioquímicas ni ecográficas). Su alta capacidad

predictiva se puede deber a: 1) uso de variables que ya conocemos que son relevantes en el ámbito de la enfermedad traumática, de tal manera que facilitamos la tarea ofreciendo un adecuado conjunto de variables a los diferentes AML; 2) integrar interacciones complejas/no lineales entre las variables que los modelos tradicionales se pierden, asemejándose al pensamiento clínico; 3) además de integrar interacciones, permite integrar variaciones en la gravedad de cada variable de manera variable y no de otorgar valores fijos a rangos dentro de una variable (fig. 5). La potencia predictiva de los AML invitaría a 3 reflexiones: 1) debería aumentar la utilización de este tipo de sistemas predictivos, especialmente en una época con crecimiento exponencial de los datos²²; 2) proactividad en la incorporación de AML a

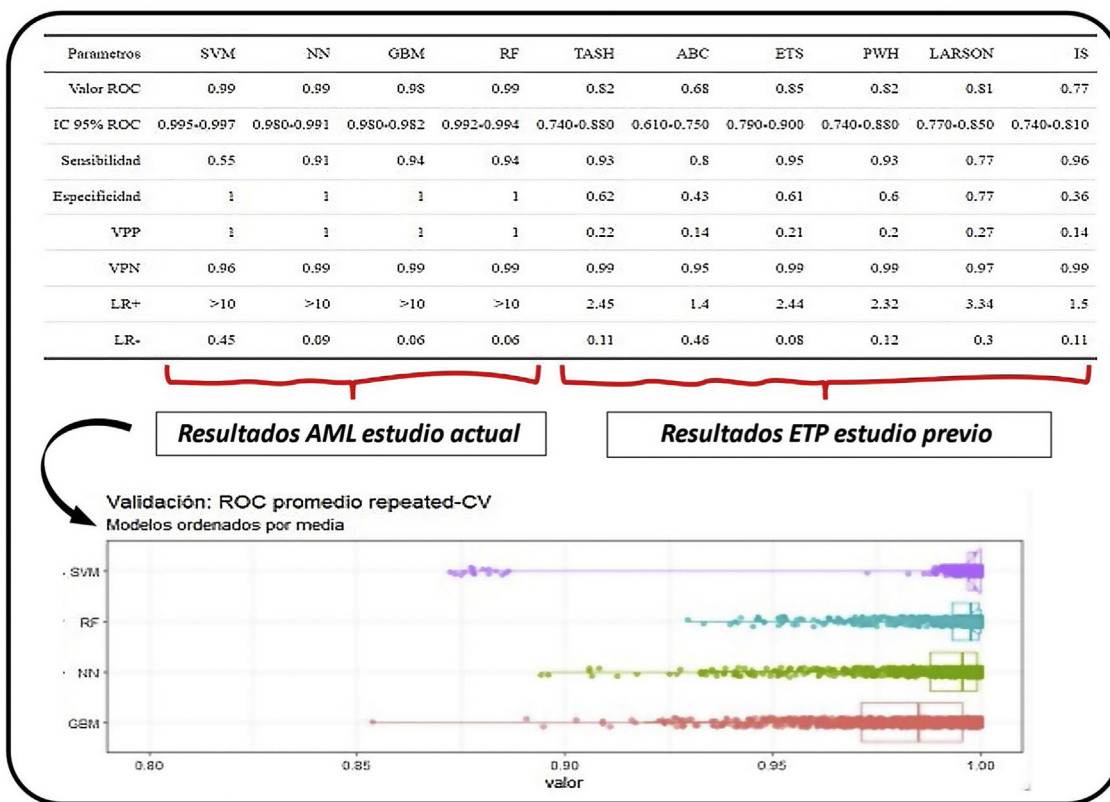


Figura 3 Resultados de AML y de ETP. En la parte superior de la imagen observamos la comparación gráfica de los resultados obtenidos por los algoritmos de *machine learning* y los *scores* predictivos tradicionales obtenidos en el estudio previo⁵. Además de la métrica ROC y del AUC con sus IC, se expresan las siguientes métricas: sensibilidad, especificidad, valores predictivos positivo y negativo, y razones de verosimilitud positiva (LR+) y negativa (LR-). En la parte inferior vemos los diferentes resultados obtenidos por los modelos de *machine learning* de donde se obtienen los estimados.

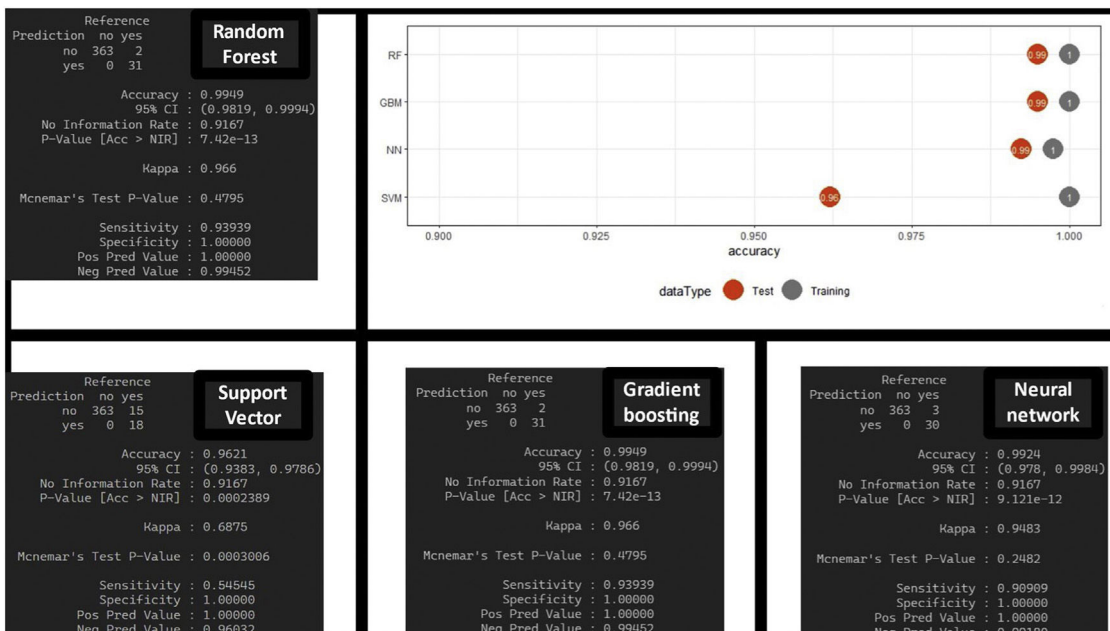


Figura 4 Matrices de confusión de cada algoritmo y resultados de aciertos en el conjunto de entrenamiento y de validación. Resultados gráficos y de la matriz de confusión de los algoritmos de *machine learning* con el conjunto de datos de test, lo que supondría una comprobación de su funcionamiento. Observamos buen rendimiento de los diferentes algoritmos.

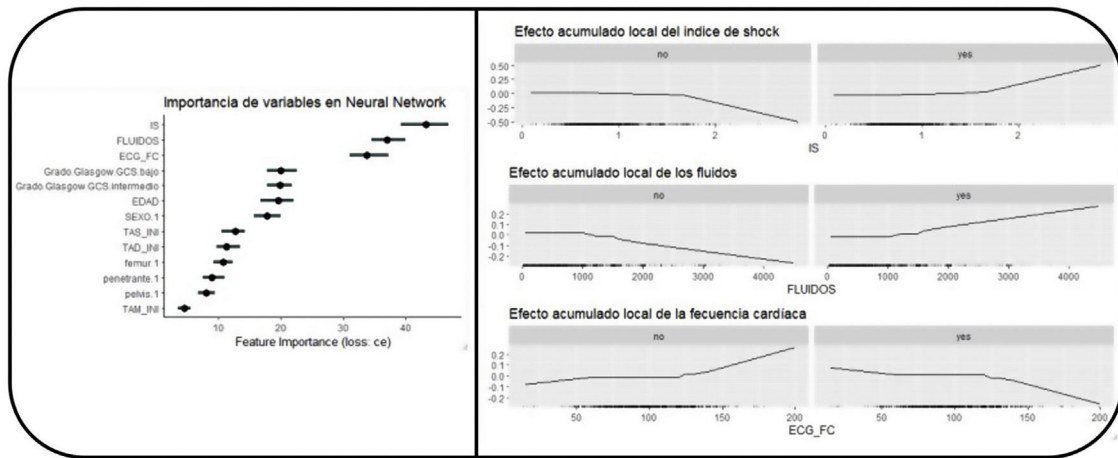


Figura 5 Importancia de las variables en el AML tipo *neural network* y la evolución no lineal de las variables. Decidimos seleccionar el algoritmo de *neural network* para explorar el comportamiento de las variables. Podemos apreciar que el comportamiento de la variable no es lineal, siendo esto uno de los fundamentos para una mayor calidad predictiva.

la práctica clínica, especialmente en enfermedades dependientes de tiempo²³, y 3) necesidad de traslado al campo prehospitalario para comenzar la predicción de la manera más precoz posible²⁴.

Limitaciones

Como limitaciones de este estudio podemos encontrar las siguientes:

- 1) Son técnicas estadísticas con un aprendizaje basado en la inducción. Esta inducción incluye inexorablemente aspectos sociosanitarios tales como: un área sanitaria en particular que tiene una interacción extrahospitalaria-hospitalaria concreta. Es decir, no podemos esperar que estos modelos puedan ser aplicados ciegamente sobre otras poblaciones si no comparten el contexto sociosanitario (se pueden incorporar los AML, pero no tan fácilmente los modelos matemáticos resultantes de los AML, ya que cada modelo sanitario tiene variables no medidas y no conocidas que influyen en el resultado).
- 2) La variable de clasificación (presencia de HM) no selecciona todo el espectro clínico en el que la activación de PTM puede ser necesaria y útil; además, se ha tenido que realizar síntesis de datos para facilitar el aprendizaje y la validación de los resultados.
- 3) Existen variables de uso frecuente en el ámbito prehospitalario que no se han utilizado o no quedaron recogidas en la base de datos de trabajo (mecanismo lesional, lactato, hemoglobina, noradrenalina prehospitalaria, resultados ecográficos prehospitalarios e intubación orotraqueal).
- 4) Se podría criticar la utilización de uno de los predictores (IS) en nuestro estudio, sin embargo, consideramos que el uso que hacemos de él no es de predictor sino de variable.
- 5) La variable «fluidos administrados prehospitalariamente» no es una variable inicial y puede tener cierto sesgo individual, si bien es cierto que

posiblemente los grandes volúmenes prehospitalarios reflejen un paciente con hemorragia crítica traumática más que una administración de fluidos laxa.

- 6) Cuando los resultados son muy buenos, siempre debemos considerar la posibilidad de sobreajuste, especialmente tras aplicar SMOTE; este hecho obligaría a desafiar al algoritmo con casos nuevos y ver su comportamiento. Si el comportamiento fuese bueno nos habríamos asegurado de que el algoritmo ha captado los patrones que reflejan la realidad, si el algoritmo tuviese un desempeño menor de lo esperado podríamos pensar que más que «captar la realidad» se ha «aprendido la muestra» (el verdadero sobreajuste).
- 7) Las métricas deseables cuando los datos son desbalanceados son: las curvas *precision-recall* o el uso del *score F1*. Se ha decidido utilizar la curva ROC para tener mayor comparabilidad conociendo que en estas situaciones es «optimista».
- 8) Se ha utilizado una base de datos pequeña y técnicas potentes de ciencias de datos, por lo que debemos ser cautos en la interpretación de los resultados, especialmente por haber aplicado SMOTE sobre el conjunto de datos total.
- 9) La confrontación de resultados del estudio previo⁵ con los del conjunto de validación no permite una adecuada comparación estadística, por lo que estos resultados únicamente plantean resultados, pero por la metodología estudiada no pueden establecer su superioridad de manera categórica.

Conclusiones

Los AML podrían superar la capacidad predictiva de los ETP y ayudar en la predicción de HM. Hacen falta estudios comparativos adecuados. Las variables más relevantes serían: hemodinámicas, de reanimación y el deterioro neurológico. Estos resultados apoyarían el desarrollo y la incorporación de este tipo de técnicas estadísticas en la práctica clínica diaria como herramientas de apoyo en la toma de decisiones.

Autoría

Todos los autores han participado en el trabajo de investigación y en la preparación del artículo. Asimismo, todos han revisado y aprobado la versión final del artículo.

1. Marcos Valiente Fernández: búsqueda bibliográfica, análisis estadístico y redacción del artículo.
2. Carlos García-Fuentes: búsqueda bibliográfica, base de datos, recogida de datos, análisis estadístico y redacción del artículo.
3. Francisco de Paula Delgado Moya: redacción del artículo.
4. Adrián Morales Marcos: metodología y redacción del artículo.
5. Hugo Fernández Hervás: metodología y redacción del artículo.
6. Jesús Abelardo Mendoza: metodología y redacción del artículo.
7. Carolina Mudarra Reche: redacción del artículo.
8. Susana Bermejo Aznárez: redacción del artículo.
9. Reyes Muñoz Calahorra: metodología y redacción del artículo.
10. Laura López García: redacción del artículo.
11. Fernando Monforte Escobar: redacción y colaboración base de datos prehospitalaria.
12. Mario Chico Fernández: redacción del artículo

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Anexo. Material adicional

Se puede consultar material adicional a este artículo en su versión electrónica disponible en [doi:10.1016/j.medin.2023.06.011](https://doi.org/10.1016/j.medin.2023.06.011).

Bibliografía

1. Norton R, Kobusingye O. Injuries. *N Engl J Med*. 2013;368:1723–30.
2. Cantle PM, Cotton BA. Prediction of massive transfusion in trauma. *Crit Care Clin*. 2017;33:71–84, <http://dx.doi.org/10.1016/j.ccc.2016.08.002>.
3. Cotton BA, Reddy N, Hatch QM, LeFebvre E, Wade CE, Kozar RA, et al. Damage control resuscitation is associated with a reduction in resuscitation volumes and improvement in survival in 390 damage control laparotomy patients. *Ann Surg*. 2011;254:598–605.
4. Mitra B, Rainer TH, Cameron PA. Predicting massive blood transfusion using clinical scores post-trauma. *Vox Sang*. 2012;102:324–30.
5. Terceros-Almanza LJ, García-Fuentes C, Bermejo-Aznárez S, Prieto del Portillo IJ, Mudarra-Reche C, Domínguez-Aguado H, et al. Prediction of massive bleeding in a prehospital setting: Validation of six scoring systems. *Med Intensiva (Engl Ed)*. 2019;43:131–8, <http://dx.doi.org/10.1016/j.medin.2017.12.005>.
6. Brockamp T, Nienaber U, Mutschler M, Wafaisade A, Peiniger S, Lefering R, et al. Predicting on-going hemorrhage and transfusion requirement after severe trauma: A validation of six scoring systems and algorithms on the TraumaRegister DGU. *Crit Care*. 2012;16:R129.
7. Greco M, Caruso PF, Cecconi M. Artificial intelligence in the intensive care unit. *Semin Respir Crit Care Med*. 2021;42:2–9, <http://dx.doi.org/10.1055/s-0040-1719037>.
8. Gutierrez G. Artificial intelligence in the intensive care unit. *Crit Care*. 2020;24:101, <http://dx.doi.org/10.1186/s13054-020-2785-y>.
9. Nunez TC, Young PP, Holcomb JB, Cotton BA. Creation, implementation, and maturation of a massive transfusion protocol for the exsanguinating trauma patient. *J Trauma*. 2010;68:1498–505.
10. SAMUR-Protección Civil. Manual de procedimientos [consultado 27 Sep 2022]. Disponible en: <https://www.madrid.es/ficheros/SAMUR/index.html>.
11. Yucel N, Lefering R, Maegele M, Vorweg M, Tjardes T, Ruchholtz S, et al. Trauma Associated Severe Hemorrhage (TASH)-Score: Probability of mass transfusion as surrogate for life threatening hemorrhage after multiple trauma. *J Trauma*. 2006;60:1228–36.
12. Maegele M, Lefering R, Wafaisade A, Theodorou P, Wutzler S, Fischer P, et al. Revalidation and update of the TASH-Score: A scoring system to predict the probability for massive transfusion as a surrogate for life-threatening haemorrhage after severe injury. *Vox Sang*. 2011;100:231–8.
13. Kalkwarf KJ, Goodman MD, Press GM, Wade CE, Cotton BA. Prehospital ABC score accurately forecasts patients who will require immediate resource utilization. *South Med J*. 2021;114:193–8. <https://doi.org/10.14423/SMJ.0000000000001236>.
14. Kuhne CA, Zettl RP, Fischbacher M, Lefering R, Ruchholtz S. Emergency Transfusion Score (ETS): A useful instrument for prediction of blood transfusion requirement in severely injured patients. *World J Surg*. 2008;32:1183–8.
15. Vandromme MJ, Griffin RL, Kerby JD, McGwin G Jr, Rue LW 3rd, Weinberg JA. Identifying risk for massive transfusion in the relatively normotensive patient: Utility of the prehospital shock index. *J Trauma*. 2011;70:384–8.
16. Terceros-Almanza LJ, García-Fuentes C, Bermejo-Aznárez S, Prieto-del Portillo IJ, Mudarra-Reche C, Sáez-de la Fuente I, et al. Prediction of massive bleeding. Shock index and modified shock index. *Med Intensiva*. 2017;41:532–8, <http://dx.doi.org/10.1016/j.medin.2016.10.016>.
17. Larson CR, White CE, Spinella PC, Jones JA, Holcomb JB, Blackburne LH, et al. Association of shock, coagulopathy, and initial vital signs with massive transfusion in combat casualties. *J Trauma*. 2010;69:S26–32.
18. Hong N, Liu C, Gao J, Han L, Chang F, Gong M, et al. State of the art of machine learning-enabled clinical decision support in intensive care units: Literature review. *JMIR Med Inform*. 2022;10:e28781, <http://dx.doi.org/10.2196/28781>.
19. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ*. 2015;350:g7594, <http://dx.doi.org/10.1136/bmj.g7594>.
20. Breiman L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statist Sci*. 2001;16:199–231.
21. Shillan D, Sterne JAC, Champneys A, Gibbison B. Use of machine learning to analyse routinely collected intensive care unit data: A systematic review. *Crit Care*. 2019;23:284, <http://dx.doi.org/10.1186/s13054-019-2564-9>.

22. Chang H, Yu JY, Yoon S, Kim T, Cha WC. Machine learning-based suggestion for critical interventions in the management of potentially severe conditioned patients in emergency department triage. *Sci Rep.* 2022;12:10537, <http://dx.doi.org/10.1038/s41598-022-14422-4>.
23. Zhao Y, Jia L, Jia R, Han H, Feng C, Li X, et al. A new time-window prediction model for traumatic hemorrhagic shock based on interpretable machine learning. *Shock.* 2022;57:48–56, <http://dx.doi.org/10.1097/SHK.0000000000001842>.
24. Lammers D, Marengo C, Morte K, Conner J, Williams J, Bax T, et al. Machine learning for military trauma: Novel massive transfusion predictive models in combat zones. *J Surg Res.* 2022;270:369–75, <http://dx.doi.org/10.1016/j.jss.2021.09.017>.