XLIII CONGRESO ANECIPN Toledo 15, 16 y 17 de Mayo de 2024

COM-043: MODELO DE APRENDIZAJE AUTOMÁTICO FRENTE A ESCALAS DE RIESGO DE MORTALIDAD COMO PREDICTORES DE EXITUS EN UCI PEDIÁTRICA

AUTORES

Vicente Mir Cerezo. Hospital Universitario y Politécnico La Fe.

RESUMEN

La UCI Pediátrica se considera un entorno desafiante por la variabilidad asociada a la edad de las constantes vitales, así como la labilidad del estado de salud de los pacientes y con un riesgo de mortalidad estimado entre el 2-6%. Los modelos de aprendizaje automático ofrecen el potencial de mejorar la predicción de mortalidad en comparación con las escalas de riesgo de mortalidad pediátricas, permitiendo una detección temprana del deterioro de la salud y aumentando las posibilidades de recuperación. Se partió de un conjunto de datos con parámetros de monitorización continua y epidemiológicos, eligiendo los datos con menor influencia del factor humano, debido a la variabilidad asociada a la interpretación de la información. Pese a trabajar con un conjunto de datos con clases desequilibradas, por el bajo porcentaje de mortalidad, se consiguió superar dicho sesgo y entrenar sin sobreajuste diversos algoritmos de aprendizaje automático. Al realizar las predicciones, se obtuvo que los modelos de redes neuronales convolucionales superaron al resto de modelos en precisión, pese a que los resultados fueron muy similares al segundo modelo con más rendimiento basado en el algoritmo de bosque aleatorio. Además, los resultados de los algoritmos de aprendizaje automático también resultaron más precisos que las escalas de riesgo de mortalidad pediátrica y el uso del sistema de alarma de los monitores.

El resultado supone una alternativa a tener en cuenta para su futura instauración en los registros médicos electrónicos o monitores. Con ello se pretende ganar poder de anticipación, servir como referencia al personal menos formado y ayudar en la toma de decisiones clínicas.

PALABRAS CLAVE: aprendizaje, automático, predicción, mortalidad.

INTRODUCCIÓN

Este proyecto se enmarca en el ámbito de la ciencia de datos aplicada a los registros médicos electrónicos (RME o EMR en inglés) en la Unidad de Cuidados Intensivos Pediátricos (UCIP). La UCIP supone un entorno complejo debido a la variabilidad presente entre los diferentes grupos de edad (de 1 mes a 14 años) y su cambio rápido del estado de salud (1). Si se suma su estado vital generalmente grave, aumenta la probabilidad de un evento adverso. La severidad, como probabilidad de evento adverso en UCIP, ronda el 9% (2). Los eventos adversos destacados en UCIP son: muerte, parada respiratoria, sedación excesiva, fallo ventilatorio, complicaciones quirúrgicas y

deterioro agudo (3). Entre ellos el riesgo de muerte oscila entre 2-6% (4,5) y tiene un impacto más dramático.

Habitualmente se utilizan monitores con alarmas de alerta temprana de las constantes vitales para vigilar y detectar precozmente empeoramientos de salud. Aunque, estos monitores suelen estar programados con parámetros estándar para la población pediátrica, lo que puede resultar en un alto porcentaje de falsos positivos o FP (6).

Por otra parte, las escalas de riesgo de mortalidad también se utilizan para predecir el pronóstico de los pacientes, pero es crucial interpretar los resultados con cautela (7). Algunas de estas escalas, como pSOFA, utilizan parámetros de laboratorio que se miden a intervalos más espaciados y son susceptibles a la variabilidad en su petición (7). La escala PEWS funciona como alarma de riesgo temprano a partir de signos vitales que incluyen: frecuencia cardiaca, frecuencia respiratoria, presión arterial, saturación de oxígeno y temperatura (8). La escala TISS-28, se calcula por carga de trabajo y se utiliza como predictor indirecto de gravedad. Investigaciones sobre el rendimiento de este tipo de escalas, sitúan el AUC-ROC promedio de 0.74 (IC95% 0.6-0.85) (9,10).

Por otra parte, no existe suficiente evidencia científica del rendimiento de la monitorización del riesgo de muerte a lo largo del episodio clínico mediante escalas de riesgo de muerte pediátrica, ya que su uso se suele restringir al ingreso y con rendimientos generalmente inferiores a los modelos predictivos (11).

Lo anterior expuesto, sumado a la ingente información clínica generada en estos servicios, ha favorecido el desarrollo de numerosas iniciativas basadas en modelos predictivos, con el fin de ganar mayor poder de detección (12). Las iniciativas han evolucionado desde el uso de modelos multivariantes de riesgos proporcionales hasta centrarse en modelos basados en Machine Learning (ML). Estos algoritmos son representaciones matemáticas que permiten realizar predicciones sobre los valores de una variable utilizando datos previos y refinando sus predicciones a través de un proceso de aprendizaje generalmente iterativo. Además, se diferencian de los modelos de riesgos proporcionales o los de regresión múltiple en que son menos dependientes del conjunto de datos, de la magnitud de variables y tienen mayor capacidad de captar relaciones no lineales complejas (13).

Los modelos de ML utilizan variables predictoras (constantes vitales y parámetros epidemiológicos), para predecir la variable respuesta y que en este caso será clasificar si muere el paciente. Con esta información el algoritmo realiza relaciones entre las variables de manera automática. Además, existe una variante de ML llamada algoritmos de aprendizaje automático (AutoML), con capacidad de aprender la estructura inherente de los datos aunque no estén etiquetados.

Los modelos resultantes de la aplicación de estos algoritmos se están diseñando para sistemas de apoyo en la toma de decisiones clínicas, principalmente en la predicción del riesgo de muerte, aunque la tendencia actual se inclina más hacia la evaluación del deterioro clínico significativo (14). Algunos ejemplos, que se han encontrado, utilizando series temporales son:

- Algoritmos de predicción basados en AutoML para desarrollar programas como el predictor de riesgo de mortalidad infantil (PROMPT), que realiza una predicción con AUC-ROC de 0.89 a las 6 h (15).

- Uso de algoritmos para la predicción de shock séptico (16).
- Predicción del tiempo de estancia de los pacientes en UCIP (17).
- Predicción del riesgo de fallo de extubación con AUC 0.75 (18).

JUSTIFICACIÓN

Es importante destacar que la mayor parte de estudios revisados se han realizado en UCI de adultos; en cambio, para la UCIP se ha encontrado poca bibliografía al respecto. Añadir que, a pesar de algunas excepciones (19,20), la validación externa (utilizar otro conjunto de datos) sigue siendo un desafío, siendo la mayoría de los estudios unicentro. En el ámbito de la interpretabilidad del modelo, se están realizando importantes contribuciones, ya que tradicionalmente los modelos de ML se percibían como "cajas negras". Se han desarrollado técnicas como la importancia de variables, la permutación de variables y la autoatención multicabeza para redes neuronales (21,22). Se pretende así trasladar conocimientos empleados en UCI a la UCIP, establecer un modelo que mejore la capacidad de generalización y sea adecuado por la importancia que asigne a cada variable.

Por consiguiente, se busca implementar el uso de modelos basados en algoritmos de aprendizaje automático que mejoren la predicción del riesgo de muerte en la UCIP a partir de los constantes vitales de monitorización continua y medición horaria (23,24). El impacto potencial de este proyecto radica en la capacidad de predecir de manera temprana la probabilidad alta del riesgo de muerte en pacientes pediátricos, con el objetivo de aumentar las posibilidades de su recuperación exitosa.

OBJETIVO PRINCIPAL

Mejorar la detección del riesgo de muerte en pacientes de la UCIP.

OBJETIVOS SECUNDARIOS

Destacan:

- Servir como información de apoyo a la toma de decisiones clínicas.
- Desarrollar una herramienta para advertir al personal menos experto.

MÉTODOS

El protocolo que se utilizó para seleccionar los mejores modelos comprende las siguientes etapas: exploración inicial, preprocesamiento de datos, entrenamiento del modelo, prueba del modelo y evaluación del rendimiento. Luego, se comparará el mejor modelo obtenido con el promedio del AUC-ROC de las escalas de riesgo de muerte pediátrica para determinar posibles diferencias significativas en su rendimiento predictivo. A continuación, se detalla cada paso del proceso.

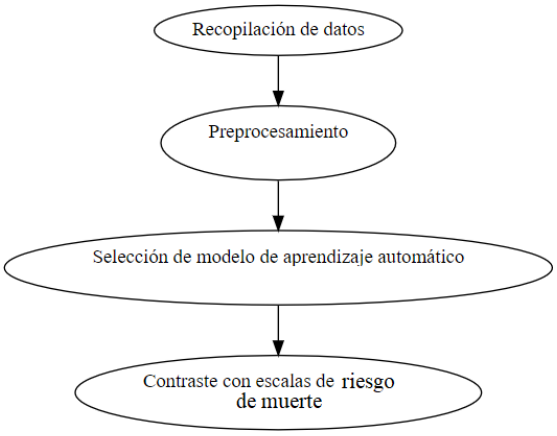

Figura 1. Diagrama del estudio. Cada óvalo incluye una etapa del protocolo para la elaboración de algoritmos de ML y sus posteriores evaluaciones.

El conjunto de datos corresponde a 90 episodios clínicos de pacientes sin identificar ingresados en UCIP. Se ha obtenido de kaggle, es de dominio público y se denomina "critically ill pediatric patients in PICU.csv"(25). Este se utilizará para desarrollar un algoritmo con el que abordar un problema de clasificación binaria relacionado con la predicción del riesgo de muerte en UCIP. El código se ha escrito con el lenguaje de programación Python en la plataforma Google Colab, por su capacidad computacional para realizar los cálculos requeridos. Cabe destacar que cuenta con librerías como Scikit-learn (26) y TensorFlow (27), desarrolladas en Python 3.10.12, y que disponen de mayor cantidad de documentación disponible. Además, Python está más extendido como lenguaje de programación, cuenta con mayor escalabilidad a otros lenguajes y aplicabilidad en programas de RME.

EXPLORACIÓN INICIAL

En este conjunto de datos se partirá de un conjunto de datos con 94678 líneas que representan una medición de cada episodio y que contiene 14 variables de tipo demográfica, clínica o signos vitales:

Tabla 1. Variables del conjunto de datos utilizado.

Variable	Tipo	Descripción
No. Patients	Cuantitativa discreta	Variable entera que identifica cada paciente pediátrico y sus mediciones asociadas.
Age	Categórica	Variable tipo objeto que indica la edad del niño en meses o años.
Weight	Cuantitativa continua	Variable decimal que mide el peso del niño en kg.

Height	Cuantitativa continua	Variable decimal que mide la altura del niño en cms.
Genero	Cualitativa	Variable tipo objeto con las categorías "hombre" (*male*) y "mujer" (*female*).
Diagnosis	Cualitativa	Variable tipo objeto que representa el diagnóstico de ingreso.
Outcome	Cualitativa dicotómica	Variable tipo objeto que clasifica el episodio clínico como "vivo" (*survived*) o "fallecido" (*dead*).
Hour event	Cuantitativa discreta	Variable tipo objeto que indica la hora del día en que se registraron las variables de cada fila del episodio en un periodo de 24 horas.
Heart Rate	Cuantitativa discreta	Variable entera que representa la frecuencia cardíaca por minuto.
Oxygen Saturation	Cuantitativa discreta	Variable entera que representa el porcentaje de oxígeno en sangre, con valores de 0 a 100%.
Respiratory Rate	Cuantitativa discreta	Variable entera que indica la frecuencia respiratoria por minuto.
Systolic Blood Pressure	Cuantitativa discreta	Variable entera que indica la presión arterial sistólica medida en mmHg.
Diastolic Blood Pressure	Cuantitativa discreta	Variable entera que indica la presión arterial diastólica medida en mmHg.
Mean Blood Pressure	Cuantitativa discreta	Variable entera que representa la medida de la presión arterial media (PAM) en mmHg, calculada según la fórmula: (2 * Presión sistólica + Presión diastólica) / 3.

**Se mantiene el nombre original de las variables en el conjunto de datos.*

El análisis del documento revela situaciones en las que se etiqueta como fallecido (dead) a pesar de estar vivo. Esto ocurre porque se clasifica la totalidad del episodio clínico y no el momento específico de la medición. Este enfoque permite a los algoritmos detectar mejor las tendencias de la evolución del episodio clínico.
Ante unas mismas constantes vitales, la evolución del episodio completo y la frecuencia de estas en cada pronóstico determinarán la clasificación.
Además, si se ciñe al criterio médico de pérdida de actividad eléctrica cardiaca y no reanimable, se pasaría de presentar un 21,69% de filas de Outcome (pronóstico) de dead a un 0.19%. Se probó previamente que este enfoque ganaba poder clasificatorio, pero se descartó por perder poder de predicción. Al desarrollar un método predictivo, se valora la capacidad de anticipar eventos de interés.

PREPROCESADO

Previo al uso de estas variables por parte de los algoritmos predictores, se realizó un preprocesado en el que se eliminaron: datos faltantes, valores atípicos, variables sin varianza significativa, con baja abundancia relativa o que no aporten información. Además, para mitigar el sesgo de selección hacia la clase mayoritaria (vivo), se realizó un sobremuestreo generando datos sintéticos para igualar clases. De las variables filtradas también se tuvo en cuenta su relevancia clínica.

ENTRENAMIENTO DEL MODELO

En este paso, el conjunto de datos se distribuyó en subconjunto de entrenamiento y de prueba. Con el subconjunto de entrenamiento se entrenaron varios algoritmos de aprendizaje automático para determinar cuál ofrece el mejor rendimiento predictivo. Tras la revisión bibliográfica, se utilizaron los algoritmos más empleados para predicción en UCIP y valorar si se puede mejorar su rendimiento con el experimento actual (16,28–30). Estos algoritmos son: bosque aleatorio (Random Forest o RF), refuerzo extremo de gradientes (Extreme Gradient Boosting o XGB), redes neuronales convolucionales (Convolutional Neural Network o CNN) y memoria a largo plazo de corto plazo (Long Short-Term Memory o LSTM).

EVALUACIÓN DEL MODELO

Lo primero que se evaluó fue si los resultados mejoraron con el sobremuestreo. Si no se han visto influenciados por el desequilibrio de clases, no debería observarse un sesgo hacia errores tipo II o falso negativo. De hecho, es preferible más errores tipo I o falso positivo cuando se pretende clasificar correctamente una clase minoritaria en un problema de clasificación binaria.
En el siguiente paso se eliminaron modelos con una área bajo la curva ROC (AUC-ROC), en el subconjunto de entrenamiento, inferior al promedio de las escalas de riesgo de mortalidad pediátrica (0.75). AUC-ROC se trata de un gráfico que muestra el rendimiento del modelo en todo el rango de valores de la tasa de verdaderos positivos o VP (eje Y) y la tasa de falsos positivos o FP (eje X); de manera que un valor cercano a 1 indican una mayor capacidad para distinguir VP y FP. Se utilizó esta métrica, pese a que obtiene mayor refuerzo bibliográfico para AUC-PR, debido a que el filtro de las escalas citadas se realizó con la métrica AUC-ROC. En cambio, los modelos con predicciones a partir del subconjunto de prueba sí se compararon utilizando AUC-PR, una métrica que evalúa la precisión y sensibilidad en problemas de desequilibrio de clases al variar el umbral de decisión del clasificador; para esta métrica, valores cercanos a 1 también indican un buen rendimiento del modelo.

Entre los modelos seleccionados, se realizó un test de Friedman. Se trata de una prueba paramétrica alternativa al ANOVA de una vía para medidas repetidas y extiende la

prueba de los rangos con signo de Wilcoxon para dos grupos. Su objetivo es determinar si existen diferencias significativas en las predicciones de ambos modelos.
También se aplicó la importancia de permutación para determinar el peso que el modelo asigna a cada variable explicativa y así determinar qué modelo es más coherente con su futura utilidad. Por último, se volvió a calcular AUC-ROC en el conjunto de prueba del modelo elegido, utilizando remuestreo mediante bootstrap para obtener los intervalos de confianza del 95% (IC 95%) y comparar si existen diferencias significativas con los IC95% de las escalas de riesgo de muerte pediátrica.

RESULTADOS

PREPROCESADO
Al centrarse en la variable respuesta Outcome (pronóstico), contiene dos clases: dead y survived (Tabla 1). La clase minoritaria es dead con 38% (25 episodios clínicos) y la clase mayoritaria es survived con un 62% (61 episodios clínicos).

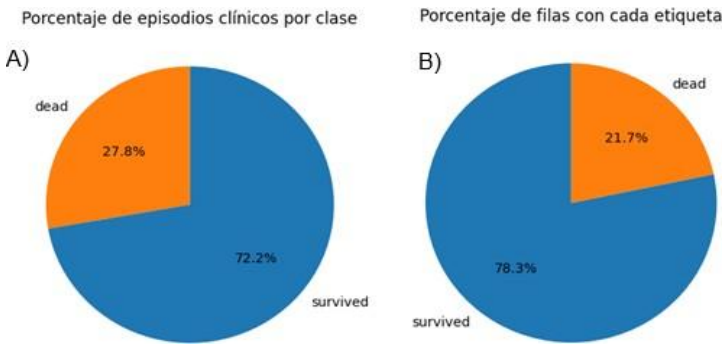

Figura 2. Diagramas circulares de la distribución de las clases Outcome. A, la sección azul representa el porcentaje de casos clínicos survived y en naranja el porcentaje de casos clínicos dead. B, diagrama circular que representa el porcentaje de filas de cada clase en el conjunto de datos y sigue el mismo código de colores.

Al explorar las variables explicativas, se realizó un test estadístico para valorar si la varianza es diferente de 0 y se obtuvo un resultado significativo del p-valor menor a 0.05 para todas las variables. Otro paso importante es conocer la correlación con la variable respuesta y por ello se realizó una matriz de correlación no lineal mediante el método de Kendall; con ello, se busca identificar las variables que más relacionadas con la variable respuesta (pronóstico). Posteriormente se determinó si las variables más relacionadas podrían producir algún tipo de sesgo en el modelo.

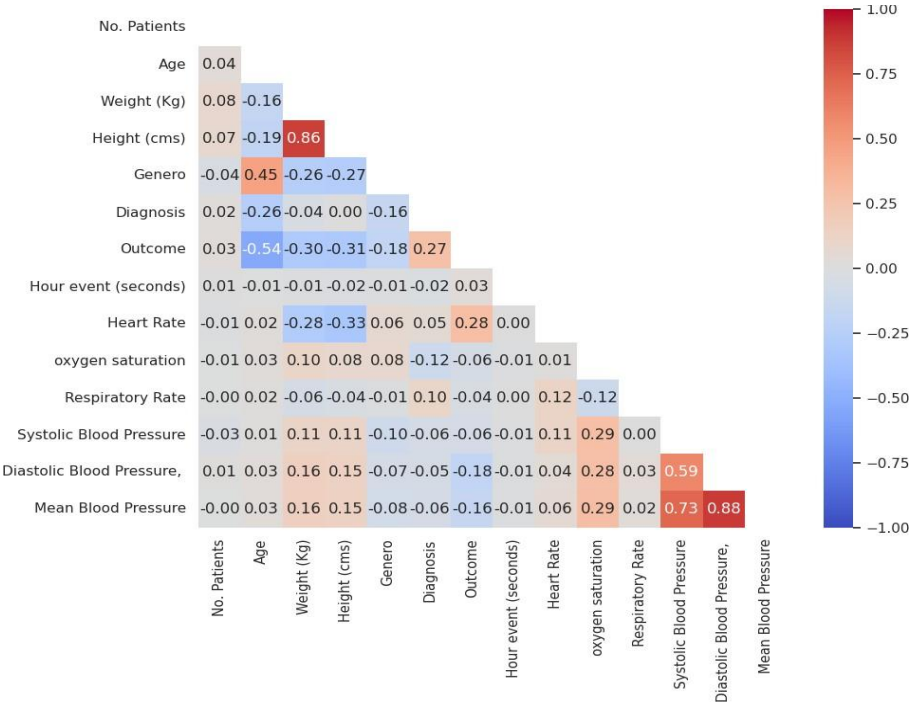

Figura 3. Matriz de correlación por coeficiente de Kendall entre todas las variables. Las correlaciones < 0.2 se muestran en azul, mientras que las correlaciones >0.2 en rojo. Además, se utiliza una escala de colores dentro de cada tonalidad: cuanto más claro el color, más cercana está la correlación a 0.2 y cuanto más oscuro, mayor es la distancia respecto a 0.2.

Se examinó la variable Diagnosis (diagnóstico), al mostrar una correlación moderada con Outcome y estar influenciada por el factor humano. Dado que una baja distribución puede afectar el peso que los modelos asigna a las variables, se analizó la distribución de sus valores:

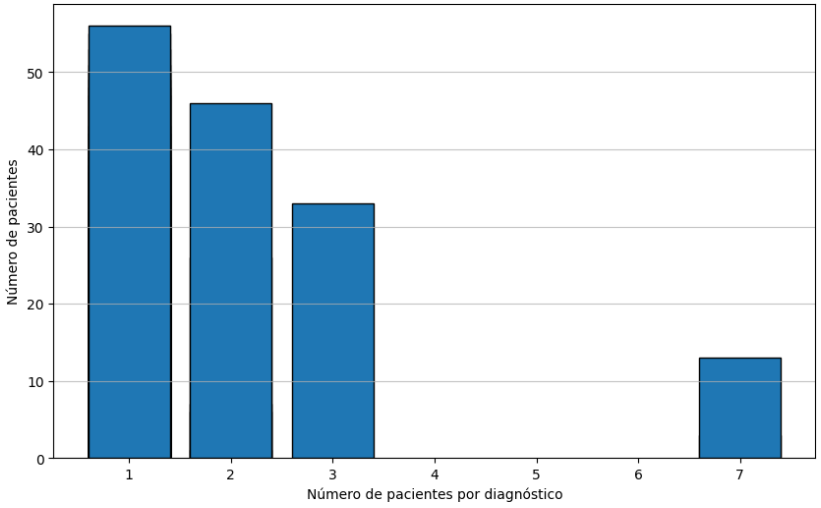

Figura 4. Diagrama de barras sobre la frecuencia del número de pacientes por diagnóstico. En el eje Y la frecuencia del número de pacientes y en el eje X el número de pacientes. Cada rectángulo azul indica la frecuencia de ese número de pacientes por diagnóstico.

Se obtiene que la mayoría de los diagnósticos aparecen 1-3 veces, por lo que hay poca representación y con distribución desigual.

EVALUACIÓN Y SELECCIÓN DEL MODELO

Tras entrenar los modelos y optimizarlos, se aplicó la validación cruzada con métrica AUC-ROC con umbral AUC-ROC 0.75, antes realizar las predicciones en el subconjunto de prueba. Como todas superaron el filtro, se realizaron las predicciones y se valoró el rendimiento AUC-PR y los VP. Los resultados se resumen en la siguiente tabla:

Tabla 2. Rendimientos de los modelos

	División episodio clínico	**División aleatoria + sobremuestreo**	**División episodio clínico + sobremuestreo**	**División episodio clínico + sobremuestreo episodio clínico**
RF	-AUC-PR: 0.971 -VP: 0.857	-AUC-PR: 1.0 -VP: 1.0	-AUC-PR: 0.973 - VP: 0.887	-AUC-PR: 0.969 - VP: 0.869
XGB	-AUC-PR: 0.926 - VP: 0.66	- AUC-PR: 0.993 - VP: 0.972	- AUC-PR: 0.967 -VP: 0.859	-AUC-PR: 0.964 -VP: 0.846
CNN	-AUC-PR: 0.949 - VP: 0.827	-AUC-PR: 0.993 - VP: 0.876	-AUC-PR: 0.967 -VP: 0.946	-AUC-PR: 0.964 -VP: 0.887
LSTM	-AUC-PR: 0.887 - VP: 0.563	-AUC-PR: 0.993 - VP: 0.876	-AUC-PR: 0.993 -VP: 0.86	-AUC-PR: 0.964 -VP: 0.887

En general, se observa que los modelos obtuvieron mejores métricas con el uso de sobremuestreo. Además, RF destacó en la mayoría de las configuraciones evaluadas, especialmente en la división aleatoria con sobremuestreo, donde alcanzó una puntuación perfecta en AUC-PR y VP. Sin embargo, CNN también mostró un rendimiento competitivo, especialmente en la división episodio clínico (mantiene las series temporales) + sobremuestreo, donde superó a RF en AUC-PR y VP. Por otro lado, XGB y LSTM mostraron resultados variables en diferentes configuraciones, pero en general, no alcanzaron el mismo nivel de rendimiento que RF y CNN. Como estos dos últimos obtuvieron resultados bastante similares, se procedió a realizar un test estadístico de Friedman para evaluar posibles diferencias significativas entre modelos. La H0 sería que no hay diferencias entre modelos y la Ha que al menos alguno de los modelos es diferente al resto. Da como resultado la existencia de diferencias significativas entre los modelos, por lo que se realizó la comparación entre pares mediante el test post-hoc de Nemenyi en que se añadió XGB como control por ser el tercer mejor modelo.

Tabla 3. Comparaciones post-hoc

Modelos	RF	CNN	XGB
RF	1.000	0.001	0.249
CNN	0.001	1.000	0.001
XGB	0.249	0.001	1.00

Al realizar un recuento mediante número de diferencias significativas (Tabla 3), el que más tenía sería CNN y que coincide con las conclusiones obtenidas al valorar el rendimiento del modelo. Pese a que el modelo CNN parece obtener mejor rendimiento, se comparó su permutación de variables con la de RF para ver cuál realiza una asignación más adecuada al problema de salud:

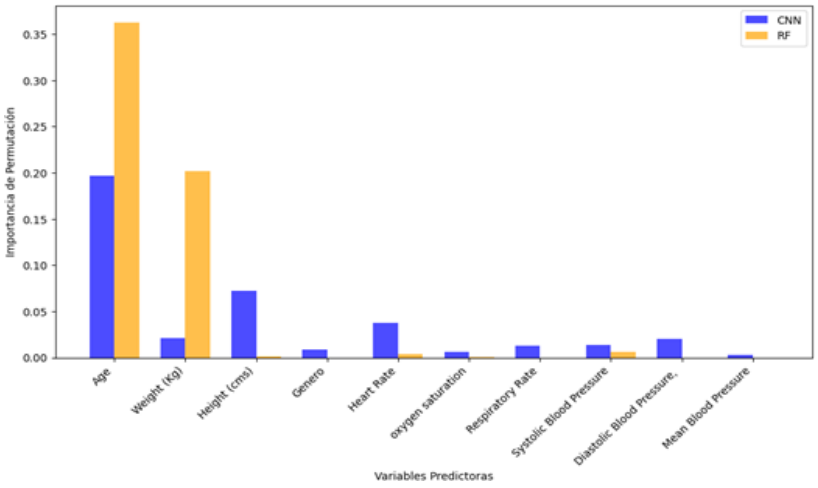

Figura 5. Análisis de la importancia de permutación para las variables predictoras en RF y CNN mediante diagrama de barras. El eje Y indica la puntuación de importancia de permutación, mientras que el eje X muestra el nombre de cada variable predictora. Los valores de la permutación de importancia obtenida en CNN se representan con barras de color azul y los de RF mediante barras amarillas.

El peso que se asigna a las variables demográficas en el caso del modelo RF es sustancialmente mayor que CNN, el cual equilibra más la distribución. De hecho, el peso que asigna RF a las constantes vitales es mínimo y CNN sí que lo contempla. Así pues, del modelo CNN se realizó bootstrap de 1000 repeticiones para obtener el IC95% del AUC-ROC de 0.984-0.987 y se obtuvo que no se solapa con el IC95% de las escalas de riesgo de muerte pediátrica (0.6-0.85). Por lo tanto, con un 95% de confianza, se puede afirmar que el verdadero valor del AUC-ROC de CNN es superior al de las escalas mencionadas.

DISCUSIÓN

Respecto a la variable respuesta, se pudo comprobar en la Figura 2 que era notablemente menor del 50% y por tanto había que tratar el problema como clases desequilibradas, lo que justificó el posterior sobremuestreo. Al analizar las variables explicativas en la Figura 3, mostraron poca correlación con la variable respuesta y ninguna es superior a 0.5 o inferior a -0.5, aunque las variables más correlacionadas positivamente fueron la frecuencia cardiaca y el diagnóstico. Esta última variable muestra una densidad de 1.37 casos por diagnósticos y una distribución desigual (baja cardinalidad). De hecho, como se puede observar en la Figura 4, que gran parte de los diagnósticos aparecen solo 1 vez, lo que podría sesgar el peso que los modelos asignen a Diagnosis; por tanto, se eliminó del conjunto de datos final. También se eliminó Hour event(seconds) y el No. Patients por no aportar información útil para el modelo predictivo. Así el conjunto de datos final mantendrá: Age, Weight(Kg), Height(cms), Diastolic Blood Pressure, Heart Rate, Mean Blood Pressure, Systolic Blood Pressure, Respiratory Rate, Oxygen Saturation, Genero.
Una vez preparado el conjunto de datos final, la división aleatoria o por episodio clínico se realizó para evaluar la importancia de la secuencia temporal en la capacidad predictiva. También se investigó si el sobremuestreo mejoraba la capacidad predictiva

en comparación con una división aleatoria. Con la combinación de estas posibilidades, se evaluó el rendimiento de los modelos.
La primera información relevante que se puede obtener de la Tabla 2 es que los modelos han obtenido mejores resultados al realizar tratar el desequilibrio de clases con sobremuestreo. Entre estos, RF mostró resultados sólidos dentro de los algoritmos tanto en datos divididos por episodio clínico (preservando la estructura temporal), como divididos aleatoriamente. XGB se utilizó como modelo supervisado, con el que tratar datos de series temporales pero que no consigue los mejores rendimientos en este tipo de algoritmos. También se observó que RF y XGB trabajan mejor con datos aleatorios que con el episodio completo, lo cual ocurre a la inversa para CNN y LSTM. Por esa razón, se eligió de los algoritmos supervisados a RF como modelo que no preserva la estructura temporal. Por parte de los algoritmos no supervisados que preservaban la estructura temporal se eligió CNN, por mejor VP que LSTM.
Aunque en principio, RF supera ligeramente a CNN en AUC-PR, posteriormente se observa un mejor desempeño de CNN para el pronóstico de muerte y de hecho resulta el modelo con más diferencias significativas al aplicar el test de Nemenyi.
Entre los modelos seleccionados, RF y CNN, en la Figura 5 se puede observar que CNN da mayor relevancia a variables dinámicas como Heart Rate (frecuencia cardíaca), Respiratory Rate (frecuencia respiratoria) y Systolic Blood Pressure (tensión arterial sistólica). Además, con diferencia significativa en AUC-ROC respecto a las escalas de riesgo de muerte pediátrica, lo que refuerza su utilidad. En conclusión, CNN ha obtenido un mejor rendimiento como modelo predictivo para este conjunto de datos, mostrando diferencias significativas en las comparaciones realizadas y siendo más sensible a constantes vitales, las cuales resultan más determinantes en el entorno sanitario.

CONCLUSIONES

En el marco AutoML, los modelos supervisados y no supervisados pueden ofrecer rendimientos altos y similares. Sobre todo, con una metodología que contemple el desequilibrio de clases. Dichos resultados son coherentes con la bibliografía revisada. Además, la automatización de procesos conlleva a que ambos modelos se puedan utilizar sin necesidad de supervisión. Ante rendimientos similares, también se tiene en cuenta aspectos prácticos. Entre ellos, la optimización es el proceso que más tiempo requiere; sin embargo, de cara a la aplicabilidad se prefiere un modelo previamente bien entrenado y optimizado. De hecho, la optimización se puede realizar como tarea de mantenimiento. Otra consideración práctica es la adecuación del peso que asigne un modelo a las variables con respecto al propósito de su desarrollo.
También es importante destacar que, más allá de mejorar la capacidad predictiva temporal, dado que pueden ocurrir intervenciones clínicas que afecten los resultados esperados, queda pendiente ganar en tamaño muestral. Se ha observado un rendimiento prometedor en este estudio, pero pendiente de aplicar en un tamaño muestral adecuado, según cálculos con EPV ≈ 1000-4000 episodios clínicos (24). Actualmente uno de los problemas a la hora de entrenar los modelos es la insuficiencia de datos debido a barreras entre centros.
Existen normativas para regular el uso de Inteligencia Artificial, como COM/2021/206 final, pero sin un uso claro en Centros Hospitalarios o de Investigación.

Se necesitan más estudios y cooperación de centros para obtener bases de datos más grandes con las que poder reentrenar los modelos y mejorar su capacidad de

predicción con validación externa. Enfermería en servicios de cuidados intensivos puede desempeñar un papel esencial y colaborar con la obtención de datos para solucionar los problemas de interés, con la función final de mejorar los cuidados de pacientes pediátricos y aumentar su probabilidad de supervivencia.

BIBLIOGRAFÍA

1. Spaeder MC, Moorman JR, Tran CA, Keim-Malpass J, Zschaebitz J V., Lake DE, et al. Predictive analytics in the pediatric intensive care unit for early identification of sepsis: capturing the context of age. Pediatr Res [Internet]. 2019 Nov;86(5):655–61. Available from: https://www.nature.com/articles/s41390-019-0518-1
2. Verlaat CW, Zegers M, Klein R, van Waardenburg D, Kuiper JW, Riedijk M, et al. Adverse Events in Pediatric Critical Care Nonsurvivors With a Low Predicted Mortality Risk: A Multicenter Case Control Study*. Pediatr Crit Care Med [Internet]. 2023 Jan 17;24(1):4–16. Available from: https://journals.lww.com/10.1097/PCC.0000000000003103
3. Agarwal S, Classen D, Larsen G, Tofil NM, Hayes LW, Sullivan JE, et al. Prevalence of adverse events in pediatric intensive care units in the United States*. Pediatr Crit Care Med [Internet]. 2010 Sep;11(5):568–78. Available from: http://journals.lww.com/00130478-201009000-00004
4. Burns JP, Sellers DE, Meyer EC, Lewis-Newby M, Truog RD. Epidemiology of Death in the PICU at Five U.S. Teaching Hospitals*. Crit Care Med [Internet]. 2014 Sep;42(9):2101–8. Available from: http://journals.lww.com/00003246-201409000-00017
5. Hajidavalu FS, Sadeghizadeh A. Mortality Rate and Risk Factors in Pediatric Intensive Care Unit of Imam Hossein Children's Hospital in Isfahan: A Prospective Cross-Sectional Study. Adv Biomed Res [Internet]. 2023;12:92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/37288019
6. Karnik A, Bonafide CP. A framework for reducing alarm fatigue on pediatric inpatient units. Hosp Pediatr [Internet]. 2015 Mar;5(3):160–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25732990
7. Shah N, Arshad A, Mazer MB, Carroll CL, Shein SL, Remy KE. The use of machine learning and artificial intelligence within pediatric critical care. Pediatr Res [Internet]. 2023 Jan 14;93(2):405–12. Available from: https://www.nature.com/articles/s41390-022-02380-6
8. Park SJ, Cho K-J, Kwon O, Park H, Lee Y, Shim WH, et al. Development and validation of a deep-learning-based pediatric early warning system: A single-center study. Biomed J [Internet]. 2022 Feb;45(1):155–68. Available from: http://www.ncbi.nlm.nih.gov/pubmed/35418352
9. Rosa RG, Roehrig C, Oliveira RP de, Maccari JG, Antônio ACP, Castro P de S, et al. Comparison of Unplanned Intensive Care Unit Readmission Scores: A Prospective Cohort Study. Salluh JI, editor. PLoS One [Internet]. 2015 Nov 23;10(11):e0143127. Available from: https://dx.plos.org/10.1371/journal.pone.0143127
10. Zhang Z, Huang X, Wang Y, Li Y, Miao H, Zhang C, et al. Performance of Three Mortality Prediction Scores and Evaluation of Important Determinants in Eight Pediatric Intensive Care Units in China. Front Pediatr [Internet]. 2020 Sep 8;8.

Available from: https://www.frontiersin.org/article/10.3389/fped.2020.00522/full
11. Hsieh MH, Hsieh MJ, Chen C-M, Hsieh C-C, Chao C-M, Lai C-C. Comparison of machine learning models for the prediction of mortality of patients with unplanned extubation in intensive care units. Sci Rep [Internet]. 2018 Nov 20;8(1):17116. Available from: https://www.nature.com/articles/s41598-018-35582-2
12. Kumar N, Akangire G, Sullivan B, Fairchild K, Sampath V. Continuous vital sign analysis for predicting and preventing neonatal diseases in the twenty-first century: big data to the forefront. Pediatr Res [Internet]. 2020 Jan 4;87(2):210–20. Available from: https://www.nature.com/articles/s41390-019-0527-0
13. Thorsen-Meyer H-C, Placido D, Kaas-Hansen BS, Nielsen AP, Lange T, Nielsen AB, et al. Discrete-time survival analysis in the critically ill: a deep learning approach using heterogeneous data. npj Digit Med [Internet]. 2022 Sep 14;5(1):142. Available from: https://www.nature.com/articles/s41746-022-00679-6
14. Mayampurath A, Jani P, Dai Y, Gibbons R, Edelson D, Churpek MM. A Vital Sign-Based Model to Predict Clinical Deterioration in Hospitalized Children*. Pediatr Crit Care Med [Internet]. 2020 Sep 8;21(9):820–6. Available from: https://journals.lww.com/10.1097/PCC.0000000000002414
15. Kim SY, Kim S, Cho J, Kim YS, Sol IS, Sung Y, et al. A deep learning model for real-time mortality prediction in critically ill children. Crit Care [Internet]. 2019 Dec 14;23(1):279. Available from: https://ccforum.biomedcentral.com/articles/10.1186/s13054-019-2561-z
16. Kamaleswaran R, Akbilgic O, Hallman MA, West AN, Davis RL, Shah SH. Applying Artificial Intelligence to Identify Physiomarkers Predicting Severe Sepsis in the PICU*. Pediatr Crit Care Med [Internet]. 2018 Oct;19(10):e495–503. Available from: https://journals.lww.com/00130478-201810000-00023
17. Rogerson CM, Heneghan JA, Kohne JG, Goodman DM, Slain KN, Cecil CA, et al. Machine learning models to predict and benchmark PICU length of stay with application to children with critical bronchiolitis. Pediatr Pulmonol [Internet]. 2023 Jun 4;58(6):1777–83. Available from: https://onlinelibrary.wiley.com/doi/10.1002/ppul.26401
18. Rooney SR, Reynolds EL, Banerjee M, Pasquali SK, Charpie JR, Gaies MG, et al. Prediction of extubation failure in the paediatric cardiac ICU using machine learning and high-frequency physiologic data. Cardiol Young [Internet]. 2022 Oct 20;32(10):1649–56. Available from: https://www.cambridge.org/core/product/identifier/S1047951121004959/type/journal_article
19. Raita Y, Camargo CA, Macias CG, Mansbach JM, Piedra PA, Porter SC, et al. Machine learning-based prediction of acute severity in infants hospitalized for bronchiolitis: a multicenter prospective study. Sci Rep [Internet]. 2020 Jul 3;10(1):10979. Available from: https://www.nature.com/articles/s41598-020-67629-8
20. Comoretto RI, Azzolina D, Amigoni A, Stoppa G, Todino F, Wolfler A, et al. Predicting Hemodynamic Failure Development in PICU Using Machine Learning Techniques. Diagnostics (Basel, Switzerland) [Internet]. 2021 Jul 20;11(7). Available from: http://www.ncbi.nlm.nih.gov/pubmed/34359385
21. Sheikhalishahi S, Bhattacharyya A, Celi LA, Osmani V. An interpretable deep learning model for time-series electronic health records: Case study of delirium

prediction in critical care. Artif Intell Med [Internet]. 2023 Oct;144:102659. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0933365723001732
22. Sun K, Marasović A. Effective Attention Sheds Light On Interpretability. 2021 May 18; Available from: http://arxiv.org/abs/2105.08855
23. Barboi C, Tzavelis A, Muhammad LN. Comparison of Severity of Illness Scores and Artificial Intelligence Models That Are Predictive of Intensive Care Unit Mortality: Meta-analysis and Review of the Literature. JMIR Med Informatics [Internet]. 2022 May 31;10(5):e35293. Available from: https://medinform.jmir.org/2022/5/e35293
24. Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Stat Methods Med Res [Internet]. 2017 Apr 19;26(2):796–808. Available from: http://journals.sagepub.com/doi/10.1177/0962280214558972
25. Ledys Izquierdo. critically ill pediatric patients in PICU [Internet]. Kaggle. Kaggle; 2020. Available from: https://www.kaggle.com/dsv/1518232
26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. 2012 Jan 2; Available from: http://arxiv.org/abs/1201.0490
27. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A system for large-scale machine learning. 2016 May 27; Available from: http://arxiv.org/abs/1605.08695
28. Chen T, Guestrin C. XGBoost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA: ACM; 2016. p. 785–94. Available from: https://dl.acm.org/doi/10.1145/2939672.2939785
29. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput [Internet]. 1997 Nov 1;9(8):1735–80. Available from: https://direct.mit.edu/neco/article/9/8/1735-1780/6109
30. Pienaar MA, Sempa JB, Luwes N, Solomon LJ. An Artificial Neural Network Model for Pediatric Mortality Prediction in Two Tertiary Pediatric Intensive Care Units in South Africa. A Development Study. Front Pediatr [Internet]. 2022 Feb 25;10. Available from: https://www.frontiersin.org/articles/10.3389/fped.2022.797080/full