Facultad de Ciencias Económicas y Empresariales
ICADE

SIMULACIÓN DE LA POLÍTICA DE TRIAJE EN URGENCIAS HOSPITALARIAS MEDIANTE GEMELOS DIGITALES Y AGENTES ENTRENADOS POR APRENDIZAJE POR REFUERZO

Autor: Beatriz Castelló Díez
Director: Eduardo Cesar Garrido Merchán

MADRID | Abril 2025

Resumen:

Este trabajo explora el uso combinado de gemelos digitales y algoritmos de aprendizaje por refuerzo (RL) como solución innovadora para mejorar el proceso de triaje hospitalario en servicios de urgencias. En un contexto de creciente presión asistencial sobre los hospitales, el triaje se convierte en una herramienta crítica para priorizar pacientes según la gravedad de su estado clínico. Sin embargo, los sistemas actuales presentan limitaciones, como la variabilidad entre profesionales y la falta de adaptación dinámica. En respuesta, se plantea el desarrollo de un gemelo digital – una réplica virtual del entorno de urgencias – sobre el cual se entrena un agente de IA utilizando técnicas avanzadas de RL.

El objetivo principal del trabajo es demostrar que un agente entrenado mediante RL, y en concreto con el algoritmo Proximal Policy Optimization (PPO), puede superar en precisión y eficiencia tanto a una política heurística basada en severidad como una política aleatoria. Para ello se implementa un entorno simulado en Python con la librería Gym, donde se modela la llegada de pacientes, la gravedad de sus síntomas, el tiempo de espera y la capacidad del hospital. En este entorno, el agente debe decidir a quién atender en cada turno con el objetivo de maximizar una recompensa diseñada para premiar la atención de pacientes graves y minimizar los tiempos de espera.

Durante el desarrollo se utilizan dos algoritmos de aprendizaje por refuerzo profundo: Deep Q-Network (DQN), un método basado en valores, y PPO, un método basado en políticas. Finalmente, se analiza la viabilidad del uso de estos sistemas en entornos reales, así como las barrearas tecnológicas, éticas y legales que deben ser abordadas para su implementación.

Palabras clave: Aprendizaje por refuerzo, gemelos digitales, triaje hospitalario, DQN, PPO, emergencias médicas, simulación clínica, recompensa, tiempo de espera, pacientes, algoritmos.

Abstract:

This paper explores the combined use of digital twins and reinforcement learning (RL) algorithms as an innovative solution to improve the hospital triage process in emergency departments. In a context of increasing care pressure on hospitals, triage becomes a critical tool to prioritize patients according to the severity of their clinical condition. However, current systems have limitations, such as variability among professionals and lack of dynamic adaptation. In response, the development of a digital twin - a virtual replica of the medical emergency environment - on which an AI agent is trained using advanced RL techniques is proposed.

The main objective of the work is to demonstrate that an agent trained using RL, and specifically with the Proximal Policy Optimization (PPO) algorithm, can outperform in accuracy and efficiency both a heuristic policy based on severity and a random policy. For this purpose, a simulated environment is implemented in Python with the Gym library, where the arrival of patients, the severity of their symptoms, the waiting time and the capacity of the hospital are modeled. In this environment, the agent must decide who to attend to in each shift with the objective of maximizing a reward designed to reward the care of serious patients and minimize waiting times.

Two deep reinforcement learning algorithms are used during the development: Deep Q-Network (DQN), a value-based method, and PPO, a policy-based method. Finally, the feasibility of using these systems in real environments is analysed, as well as the technological, ethical and legal barriers that must be addressed for their implementation.

Keywords: Reinforcement learning, digital twins, hospital triage, DQN, PPO, medical emergencies, clinical simulation, reward, waiting time, patients, algorithms.

Índice de contenidos

Índice de ilustraciones

Índice de tablas

Índice de ecuaciones

1. Introducción y motivación del artículo

1.1 Contexto

El sistema sanitario enfrenta retos cada vez más complejos debido a la creciente demanda de servicios de urgencias, el envejecimiento poblacional, las crisis sanitarias globales y la limitada disponibilidad de recursos humanos y materiales. En España, se atienden anualmente más de 32 millones de urgencias de atención primaria, con una frecuentación que ha aumentado un 13% desde 2012, según el Ministerio de Sanidad (2023). Este aumento sostenido pone en evidencia las carencias estructurales y organizativas que dificultan la gestión eficiente de los servicios de urgencias.

A esta presión se suma la necesidad de tomar decisiones rápidas y efectivas en contextos donde los recursos son insuficientes y el personal médico está sobrecargado. La pandemia del COVID-19 subrayó estas vulnerabilidades, con tasas de hospitalización que llegaron a superar las 50 por cada 1.000 habitantes en comunidades como Madrid y Cataluña (Ministerio de Sanidad, 2023). La capacidad de adaptación de los sistemas sanitarios fue puesta a prueba, evidenciando la importancia de contar con herramientas innovadoras que optimicen la toma de decisiones y la asignación de recursos.

En este contexto, el triaje hospitalario surge como un proceso crítico para priorizar la atención de los pacientes según la gravedad de sus condiciones. Sin embargo, los sistemas tradicionales de triaje, como el Sistema de Triaje de Manchester (MTS), enfrentan limitaciones significativas, como la subjetividad en las evaluaciones, la falta de uniformidad en los criterios entre hospitales y países, y el estrés que presentan los profesionales sanitarios durante su aplicación (Soler et al., s.f.). Estas deficiencias motivan a la exploración de soluciones tecnológicas avanzadas para mejorar la precisión y eficiencia del triaje.

1.2 El problema del triaje en urgencias hospitalarias

El triaje en los servicios de urgencias hospitalarias es un sistema crítico para gestionar eficazmente el flujo de pacientes y optimizar el uso de recursos médicos en situaciones donde la demanda excede la capacidad. Este proceso se centra en evaluar y clasificar en el menor tiempo posible a los pacientes según sea la gravedad de sus condiciones, asegurando que aquellos que tengan necesidades más urgentes reciban atención

prioritaria (Clínica Universidad de Navarra, s.f.). No obstante, hay numerosos retos y dificultades vinculados al triaje que demandan un cuidado continuo y mejoras fundamentadas en pruebas científicas y tecnológicas.

Uno de los principales problemas que enfrenta el triaje en los servicios de urgencia es la sobrecarga. Conforme las emergencias médicas se tornan más habituales y complejas, los centros hospitalarios sufren un incremento en el número de pacientes, lo que podría llevar a una saturación del sistema. Esta aglomeración conduce a largos periodos de espera, no solo para los pacientes menos críticos, sino también para aquellos que requieren asistencia inmediata si los recursos se encuentran mal repartidos (Lopera Betancur et al., 2022).

1.3 Motivación hacia la solución

El uso de tecnologías innovadoras como los gemelos digitales y los algoritmos de aprendizaje por refuerzo ofrece una oportunidad única para transformar el triaje hospitalario. Los gemelos digitales, que son réplicas virtuales de sistemas reales, permiten simular y analizar escenarios clínicos en tiempo real, optimizando la toma de decisiones y la asignación de recursos (Foundtech, 2022). En este sentido, estas herramientas no solo pueden mejorar la eficiencia del triaje, sino también proporcionar un entorno seguro para entrenar a profesionales sanitarios en la toma de decisiones críticas (VIU, 2023).

Por su parte, el aprendizaje por refuerzo profundo permite entrenar agentes de inteligencia artificial para tomar decisiones basadas en datos históricos y patrones clínicos, minimizando la subjetividad y mejorando la precisión en la clasificación de pacientes. Estos agentes pueden optimizar criterios como la reducción del tiempo de espera colectivo o la minimización del riesgo individual, adaptándose dinámicamente a las condiciones del entorno hospitalario (Estebaranz Santamaría, 2017).

A lo largo de este trabajo, se abordará en primer lugar un repaso del estado del arte del aprendizaje por refuerzo en medicina, analizando las principales investigaciones y aplicaciones existentes. A continuación, se definirá el alcance del TFG, incluyendo los objetivos, hipótesis y limitaciones del estudio. Posteriormente, se presentará la metodología utilizada para el desarrollo del simulador y el entrenamiento de los agentes,

seguida por la descripción de los experimentos realizados y sus resultados. Finalmente, se expondrán las conclusiones principales y se propondrán líneas de investigación futura.

2. Estado del Arte

Para comprender el valor añadido que puede aportar el aprendizaje por refuerzo al triaje hospitalario, es fundamental analizar primero su evolución y aplicaciones dentro del ámbito sanitario. En esta sección se revisa la literatura más relevante sobre su uso en medicina, se presentan casos prácticos que ilustran su impacto en distintos contextos clínicos, y se comparan sus capacidades con otros métodos de simulación aplicados al triaje.

2.1 Revisión de la literatura sobre el aprendizaje por refuerzo en el sector Sanitario

El aprendizaje por refuerzo se ha consolidado en los últimos años como una técnica prometedora para apoyar la toma de decisiones clínicas complejas. A diferencia del aprendizaje supervisado tradicional, el RL entrena a un agente que aprende políticas óptimas mediante interacción secuencial con un entorno, recibiendo recompensas por decisiones acertadas. Esta capacidad de optimizar estrategias a largo plazo y personalizar acciones en función de datos de pacientes ha motivado múltiples investigaciones recientes en medicina. De hecho, varios estudios destacan el potencial transformador del RL en salud, especialmente para decisiones médicas secuenciales bajo incertidumbre (Jayaraman et al., 2024).

Una línea destacada de trabajo es el uso de RL para optimizar tratamientos médicos personalizados. Numerosos estudios han explorado cómo un agente de RL puede aprender políticas terapéuticas adaptadas al estado cambiante de un paciente, mejorando resultados. Por ejemplo, se han desarrollado agentes que ajustan dosificaciones de fármacos de forma dinámica: optimización de sedantes en UCI, regímenes individualizados de quimioterapia oncológica, dosificación de insulina en diabetes tipo 1 usando simuladores aprobados por la FDA, o combinaciones óptimas de medicamentos para mitigar síntomas de Parkinson (Jayaraman et al., 2024). En cada caso, el RL busca secuencias de dosificación que maximicen métricas de salud del paciente (p. ej., estabilidad glicémica en diabetes o control sintomático en Parkinson). Asimismo, en oncología preventiva se diseñaron políticas de cribado de cáncer de mama basadas en riesgo individual, logrando mejorar la detección temprana en simulaciones respecto a pautas estáticas (Yala et al., 2021). También, en 2023, un

equipo dirigido por Harald Kittler en la MedUni de Vienna desarrolló un modelo de RL para diagnóstico dermatológico que, al incorporar criterios clínicos en su función de recompensa, mejoró la precisión diagnosticada global en un 12%, imitando decisiones expertas y evitando errores graves mejor que los algoritmos tradicionales (News-Medical, 2023).

Otro campo emergente es la aplicación de RL a procesos operativos y de gestión sanitaria, más allá de la decisión clínica directa. Por ejemplo, en el campo de la ventilación mecánica, Peine et al. (2021) validaron un algoritmo de RL (VentAI) capaz de ajustar dinámicamente los parámetros del ventilador (volumen tidal, PEEP, FiO2) en pacientes críticos, optimizando la oxigenación y reduciendo el riesgo de lesión pulmonar. En análisis retrospectivos sobre bases de datos UCI, el agente superó la estrategia estándar de médicos, asociándose a una mayor supervivencia a 90 días (Peine et al., 2021). Esto evidencia que un agente puede aprender políticas de control individuales de cada paciente, algo muy valioso dada la heterogeneidad de la respuesta a la ventilación. También en cirugía robótica se investiga el RL para dotar de autonomía a robots quirúrgicos. Algunos trabajos han entrenado agentes mediante Deep Q-Learning o métodos de políticas (p. ej., PPO) en simuladores quirúrgicos avanzados, logrando que un robot aprenda a optimizar sus trayectorias instrumentales durante una operación (Verma et al., 2022). Aunque estas aplicaciones robóticas están en fase experimental, indican el potencial del RL para mejorar la precisión y eficiencia intraoperatoria en el futuro.

Pese a los progresos, la literatura también enfatiza desafíos importantes al aplicar RL en medicina. Uno de los mayores retos es la seguridad y validez clínica. A diferencia de entornos de juego o simulación, en salud no es viable un enfoque de prueba y error directamente sobre pacientes. Por ello, muchos estudios entrenan y evalúan los agentes con datos históricos (aprendizaje fuera de línea) y simulaciones antes de considerar el despliegue real (Jayaraman et al., 2024). Esto limita la capacidad de los agentes para adaptarse a situaciones imprevistas y supone un desafío de generalización. Asociado a lo anterior, la escasez de datos de calidad y sesgos en los conjuntos clínicos pueden afectar al desempeño del RL; si el agente aprende de datos no representativos, sus

recomendaciones podrían ser subóptimas o injustas para ciertos grupos de pacientes (Luo et al., 2022).

Otro tema crítico es la interpretabilidad. Los modelos de RL profundo suelen ser cajas negras, lo que dificulta la confianza de los médicos en sus sugerencias (Topol, 2019). Se están explorando técnicas para extraer reglas comprensibles o incorporar restricciones éticas en la recompensa, pero sigue siendo indispensable asegurar que las decisiones del agente sean explicables y alineadas con la práctica clínica (Gaube et al., 2021). Por último, existen barreras regulatorias y éticas: la introducción de un "agente inteligente" en procesos asistenciales plantea cuestiones de responsabilidad ante errores y requiere marcos regulatorios claros (Esteva et al., 2021).

2.2 Ejemplos de aplicaciones existentes en medicina y su impacto

Aunque muchas investigaciones de RL en medicina se encuentran en fases de validación in silico o pruebas piloto, en los últimos años han surgido aplicaciones concretas que ilustran el potencial impacto de esta tecnología en entornos sanitarios reales.

Se han desarrollado asistentes que utilizan RL para guiar el proceso diagnóstico de forma interactiva. Un ejemplo reciente es el sistema de inquiry médica propuesto por Zou et al. (2024), que emplea un modelo de aprendizaje por refuerzo actor-crítico profundo para simular la lógica de investigación clínica de un médico. Este agente entrena su política realizando preguntas secuenciales al paciente (síntomas, antecedentes) y solicitando pruebas en varias etapas, recibiendo recompensas al acercarse al diagnóstico correcto con el mínimo de indagaciones (Zou et al., 2024). En validaciones con datos reales de consultas, el asistente alcanzó alta precisión diagnóstica (curvas ROC con AUC ~ 0,95) en escenarios de emergencias y pediatría. Además, en simulaciones colaborativas logró reducir casi la mitad el número de preguntas que un médico necesita hacer (un 46% del interrogatorio habitual) manteniendo una exactitud incluso mayor (AUC ~ 0,97) (Zou et al., 2024). Esto indica que el agente aprendió a preguntar de forma muy eficiente, concentrándose en la información más relevante para el caso. En conclusión, el algoritmo RL replicó la estrategia de un médico experimentado: primero recabar síntomas clave, luego

examinar o pedir pruebas específicas, y finalmente proponer un diagnóstico cuando la evidencia es suficiente. Este caso demuestra el potencial de RL para agilizar el proceso diagnóstico asistiendo a los clínicos en la toma de decisiones informadas, especialmente en entornos donde el tiempo es crítico.

Otro ejemplo, es la gestión de pacientes con sepsis en UCI, que ha servido como banco de pruebas para la aplicación de RL en decisiones terapéuticas. En 2023, se desarrolló un agente de aprendizaje por refuerzo profundo de tipo value-based integrado con conocimiento experto para recomendar en tiempo real las acciones óptimas en el manejo del paciente séptico (Choi et al., 2024). El modelo fue entrenado con miles de historiales de pacientes, aprendiendo una política que indicaba en cada momento si administrar fluidos, vasopresores, antibióticos u otras intervenciones, con el objetivo de maximizar la probabilidad de supervivencia. Este agente, entrenado mediante un algoritmo similar a Deep Q-Network, pero con ajustes para incorporar criterios clínicos humanos, logró reproducir decisiones congruentes con las de médicos intensivistas, a la vez que descubría secuencias de tratamiento eficaces que podían no ser evidentes de forma tradicional (Choi et al., 2024). Los resultados, publicados en NPJ Digital Medicine, mostraron que la política de RL coincidía frecuentemente con las decisiones tomadas por expertos y sugería optimizaciones en el uso de vasopresores y fluidos, evidenciando la capacidad del RL para encontrar trade-offs óptimos en situaciones de manejo complejo de la sepsis (Wu et al., 2023). Por ejemplo, el agente tendía a iniciar vasopresores ligeramente antes en ciertos pacientes inestables, anticipando hipotensiones, lo que se tradujo en mejores métricas de perfusión en las simulaciones. Asimismo, al incorporar “penalizaciones” por exceso de líquidos, el algoritmo aprendió a evitar sobrecargas, sugiriendo volúmenes más ajustados que los propuestos por protocolos genéricos. En la cohorte de validación MIMIC-III, la política del agente habría conseguido una tasa de supervivencia del 98% de los pacientes (frente a 95% real) bajo ciertas condiciones iniciales. Aunque esta cifra debe interpretarse con cautela, ilustra el techo de mejora potencial identificado por la IA (Wu et al., 2023).

En España, se están implementando sistemas de triaje digital apoyados por algoritmos de inteligencia artificial que aprenden de datos históricos para priorizar pacientes. Un ejemplo es el Hospital QuirónSalud Clideba, que en 2023 anunció la optimización de su

circuito de urgencias médicas mediante un servicio de “urgencia digital” integrado en la atención prehospitalaria (Quirónsalud, 2023). El sistema se basa en una plataforma de telemedicina y una aplicación móvil por la que el paciente, al tener una urgencia, realiza una consulta inicial remota. Detrás de esta consulta, un algoritmo inteligente (entrenado con machine learning y refinado con aproximaciones de aprendizaje por refuerzo) evalúa los síntomas reportados y la información clínica del paciente en tiempo real. Según el Dr. Manuel Moreno, coordinador de urgencias, este sistema logro resolver alrededor del 60% de las consultas urgentes de forma telemática, dando recomendaciones o pautas sin necesidad de acudir físicamente al hospital. Esto supuso una importante disminución de la carga asistencial presencial y de la posible saturación de la sala de espera, permitiendo que los casos leves se atendieran virtualmente con rapidez mientras los recursos de urgencias se concentraban en pacientes de mayor gravedad. Para los pacientes que sí tenían que acudir, el sistema ya había realizado un triaje preliminar durante la teleconsulta, de modo que llegaban con una prioridad asignada y, en algunos casos, con pruebas diagnosticas iniciales indicadas. Este triaje digital está basado en la IA y se retroalimenta con cada nueva atención, ajustando sus recomendaciones futuras en función de los resultados (p.ej., si cierta combinación de síntomas derivó en ingreso, el sistema “aprende” a asignar mayor prioridad a esos síntomas en consultas posteriores) (Quirónsalud, 2023).

Por último, más allá de los algoritmos aislados, una tendencia creciente es la integración de RL con gemelos digitales. Un ejemplo pionero es el del Hospital Universitario Ramón y Cajal de Madrid, que ha empleado un gemelo digital para mejorar la gestión de su servicio de radiología (Norlean, 2021). En concreto, se desarrolló un modelo virtual de la sala de TAC (Tomografía Axial Computarizada) que reproducía el flujo de pacientes, la programación de citas, disponibilidad de personal y equipamiento. Sobre este entorno simulado, un algoritmo de aprendizaje exploró distintas políticas de asignación de turnos y priorización de estudios urgentes. Como resultado, el hospital identificó oportunidades de mejora que en la práctica real serían difíciles de anticipar. Tras implementar recomendaciones derivadas del gemelo digital, se reportó una reducción significativa en los tiempos de espera para pacientes ambulatorios de TAC, al evitar cuellos de botella en la agenda (Villalba, 2024). Directivos del proyecto

mencionaron que el tiempo promedio de espera para un TAC no urgente bajó de semanas a días, acercándose al estándar recomendado. Además, se optimizó el uso del escáner, disminuyendo horas ociosas y recortando costes asociados a horas extra del personal técnico. Tecnológicamente, en este proyecto el RL jugó el papel de motor de optimización; el agente probó virtualmente miles de combinaciones de horarios y protocolos de triaje interno. La política óptima aprendida por el agente luego sirvió de guía para rediseñar el proceso real. Este caso demuestra cómo un gemelo digital combinado con aprendizaje por refuerzo permite tomar decisiones informadas de gestión sanitaria, con impacto directo en la calidad asistencial. De forma análoga, otros hospitales como el Mater Private de Dublín, también han utilizado gemelos digitales para optimizar recursos; en este centro, la simulación con RL en el departamento de radiología logró disminuir aproximadamente 30 minutos el tiempo de espera por paciente y reducir el coste de personal (Siemens Healthineers, 2019). En resumen, los gemelos digitales hospitalarios habilitan un entrenamiento y prueba de estrategias en un "mundo virtual", cuyos hallazgos luego se traducen en mejoras reales en eficiencia, lo que resulta especialmente útil en áreas complejas como urgencias, gestión de camas o planificación de cirugías.

2.3 Comparación con otros enfoques de simulación en triaje hospitalario

El triaje hospitalario se ha estudiado mediante diversas técnicas de simulación y modelos alternativos al aprendizaje por refuerzo, cada una con sus fortalezas. Entre otras, se encuentran la simulación por eventos discretos, modelo basado en agentes, lógica difusa y redes bayesianas.

La simulación de eventos discretos (SED), modela el flujo de pacientes y recursos en el servicio de urgencias como una secuencia cronológica de eventos (llegadas de pacientes, inicios de atención, altas, etc.), permitiendo evaluar el impacto de cambios operativos en métricas como tiempos de espera o saturación. Por ejemplo, Peng et al (2020) construyeron un modelo de eventos discretos para simular el proceso de triaje y atención en urgencias, validándolo con datos reales de flujo de pacientes. Sus experimentos mostraron que introducir un médico en la etapa de triaje reducía el tiempo de estancia promedio en un 34% y el tiempo de espera en un 49%, aliviando significativamente la congestión (Peng et al., 2020). Este enfoque facilita el análisis de

escenarios hipotéticos (e.g., agregar personal, reorganizar prioridades...) antes de aplicarlos a la realidad, aunque depende de supuestos predefinidos (distribuciones de tiempos de servicio, rutas de pacientes) y no adapta automáticamente las decisiones como haría un algoritmo de RL.

Por otro lado, en el **modelado basado en agentes (MBA)** se simulan individuos ("agentes") con comportamientos autónomos dentro del entorno del hospital. En triaje, los agentes típicos pueden ser pacientes (con distintos tipos de gravedad y tiempos de llegada) y personal médico (médicos, enfermeros de triaje), cuyas interacciones emergen en dinámicas complejas. Terning et al. (2022) desarrollaron un modelo híbrido agente-basado para una urgencia hospitalaria bajo condiciones de pandemia COVID-19, donde los pacientes y salas de atención fueron representados como agentes. Su modelo, calibrado y validado con datos reales, pudo reproducir con fidelidad la ocupación de pacientes y probar políticas de gestión durante la pandemia (Terning et al., 2022). Este tipo de simulación basada en agentes permite incorporar heterogeneidad en pacientes (por ejemplo, distintas patologías y su evolución) y reglas de decisión descentralizadas; además, puede evaluar estrategias de triaje adaptativas (como ajustar criterios de priorización según la afluencia). Sin embargo, el MBA conlleva mayor complejidad computacional y de modelado, y sus resultados pueden ser sensibles a las reglas de comportamiento definidas para cada agente. A diferencia de un enfoque de RL, donde el agente "aprende" la mejor política, en MBA el comportamiento suele programarse a priori o seguir heurísticas basadas en la experiencia clínica.

La toma de decisiones difusa es útil en triaje debido a la naturaleza imprecisa y lingüística de muchas evaluaciones clínicas iniciales (síntomas vagos, dolor "moderado" vs "severo", etc.). Los **sistemas de lógica difusa** aplican reglas del tipo "Si X entonces Y" con variables borrosas (temperatura alta, dolor medio) en lugar de umbrales estrictos, produciendo una clasificación de prioridad más flexible. Galo et al (2022) propusieron un sistema difuso para apoyar el triaje de pacientes sospechosos de COVID-19 en Brasil, integrando síntomas y signos vitales en un modelo de inferencia. Tras un piloto con datos reales, el modelo mostró resultados concordantes con las evaluaciones humanas de severidad, sugiriendo que podría agilizar el proceso de triaje y reducir el número de personal expuesto sin perder precisión en la priorización (Galo et

al., 2022). La lógica difusa ofrece interpretabilidad, ya que sus reglas pueden entenderse clínicamente y maneja incertidumbre de forma explícita, lo que es valioso para completar o validar decisiones de triaje. No obstante, el diseño de las funciones de membresía y reglas difusas depende del conocimiento de expertos; a diferencia de RL, un sistema no "aprende" automáticamente de datos, sino que requiere ajuste manual o métodos de optimización para refinar sus reglas.

Por último, las **redes bayesianas** (o modelos de creencia bayesianos) son grafos probabilísticos que representan variables clínicas (síntomas, resultados de pruebas, etc.) y sus dependencias condicionales. En triaje, una red bayesiana puede inferir la probabilidad de cierto desenlace (por ejemplo, necesidad de ingreso, riesgo vital) dada la información inicial del paciente. Su fortaleza radica en el manejo formal de la incertidumbre y la posibilidad de actualizar creencias conforme llega nueva información. Un ejemplo reciente lo brinda Sullivan et al. (2023), quienes desarrollaron una red bayesiana para predecir la necesidad de transmisión sanguínea en niños politraumatizados, usando datos disponibles al ingreso por triaje. El modelo integro 14 variables, entre ellas signos vitales, mecanismo lesional, y alcanzó una gran capacidad predictiva (AUC aprox. 0,92) para anticipar hemorragias que requieren transfusión urgente. Este desempeño superó al de escalas clínicas tradicionales, demostrando como un enfoque bayesiano puede apoyar al personal de triaje en la detección precoz de pacientes críticos (Sullivan et al., 2023) . Las redes bayesianas, al igual que la lógica difusa, ofrecen transparencia en sus razonamientos (es posible rastrear qué factores contribuyen a una cierta probabilidad) y pueden incorporarse en sistemas de soporte a la decisión clínica en tiempo real. Además, requieren una cuidadosa elaboración del grafo casual y entrenamiento/validación con datos históricos; además, como método estático no ajustan su estructura por sí solos una vez desplegados.

En resumen, los enfoques tradicionales de simulación y modelados (SED y MBA) permiten recrear el entorno de triaje para probar intervenciones organizativas, mientras que las técnicas de inteligencia artificial simbólica o probabilística (lógica difusa y redes bayesianas) aportan sistemas de apoyo que manejan la incertidumbre inherente al triaje. A diferencia del RL que aprende políticas optimas mediante iteración y retroalimentación, estos métodos alternativos generalmente no aprenden de la

experiencia directa, sino que dependen de reglas predefinidas, distribuciones conocidas o entrenamiento supervisado con datos históricos.

TABLA 1

Autor	Título	Año
Choi et al.	Deep reinforcement learning extracts the optimal sepsis treatment policy from treatment records	2024
Foundtech	Servicios y soluciones de gemelos digitales	2022
Galo et al.	A fuzzy approach to support decision-making in the triage process for suspected COVID-19 patients in Brazil. Applied Soft Computi	2022
Gaube et al.	Do as AI say: susceptibility in deployment of clinical decision-aids	2021
Jayaraman et al.	A primer on Reinforcement learning in Medicine for clinicians	2024
Luo et al.	DTR-Bench: An in silico Environment and Benchmark Platform for Reinforcement Learning Based Dynamic Treatment Regime	2024
News-Medical	Reinforcement learning improves performance of AI-based skin cancer diagnosis	2023
Norlean	Inteligencia Artificial en Sanidad | Gemelos Digitales	2021
Peine et al.	The co-constitution of ageing and technology – a model and agenda	2021
Peng et al.	Evaluation of physician in triage impact on overcrowding in emergency department using discrete-event simulation	2020
Quirónsalud	Urgencia Digital: La tecnología al servicio de las necesidades sanitarias	2023
Siemens Healthineers	El valor del Gemelo Digital en el sector sanitario. Hospitecnia	2019
Sullivan et al.	Development and Validation of a Bayesian Belief Network Predicting the Probability of Blood Transfusion after Pediatric Injury	2023
Terning et al.	Modeling Patient Flow in an Emergency Department under COVID-19 Pandemic Conditions: A Hybrid Modeling Approach	2022
Villalba	Gemelos digitales e inteligencia artificial: el futuro de la sanidad ya está aquí	2024
Wu et al.	A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis. npj Digital Medicine	2023
Yala et al.	Multi-Institutional validation of a Mammography-Based breast cancer risk model	2021
Zou et al.	AI-Driven Diagnostic Assistance in Medical Inquiry: Reinforcement Learning Algorithm Development and Validation	2024

Bibliografía de la literatura clasificada por autor, título y año

Fuente: Elaboración propia

3. Alcance del TFG

3.1 Objetivos del trabajo: mejora de la precisión y rapidez en el triaje

El principal objetivo de este proyecto es mejorar la precisión y rapidez en el proceso de triaje hospitalario mediante el uso de gemelos digitales y técnicas avanzadas de aprendizaje por refuerzo. A continuación, se desglosan los objetivos específicos:

1. Desarrollar un modelo de gemelo digitales que replique el proceso de triaje hospitalario, permitiendo la simulación de diferentes escenarios clínicos y patologías.
2. Entrenar agentes de IA mediante aprendizaje por refuerzo permitiendo que se puedan tomar decisiones de triaje de forma autónoma y eficiente, sobrepasando las limitaciones de la toma de decisiones humanas. Estos agentes se entrenan para optimizar criterios como la minimización del tiempo de espera colectivo o la reducción del riesgo individual.
3. Comparar la precisión y eficiencia del triaje realizado por un agente entrenado mediante aprendizaje por refuerzo frente a un agente aleatorio en un entorno simulado, evaluando métricas clave como el tiempo de respuesta y la precisión en la clasificación de pacientes.
4. Explorar las potenciales mejoras que el uso de gemelos digitales y la IA basada en aprendizaje por refuerzo pueden ofrecer al proceso de formación médica, proporcionando a los médicos en formación un entorno seguro para experimentar con casos clínicos simulados
5. Evaluar la viabilidad tecnológica y ética de la implementación de sistemas de IA y gemelos digitales en entornos hospitalarios reales, identificando cuales son las limitaciones que presenta las tecnológicas actuales y las barreras éticas que se deben abordar para su uso generalizado.

3.2 Hipótesis de partida: el desempeño de un agente entrenado puede superar al de un agente aleatorio

La hipótesis principal de este trabajo, H1, es que, el desempeño del agente entrenado mediante aprendizaje por refuerzo en un entorno simulado de triaje hospitalario, será

significativamente mejor que el de un agente aleatorio. Se espera que el agente basado en aprendizaje por refuerzo optimice la clasificación de pacientes, reduciendo el tiempo de respuesta y mejorando la precisión de la toma de decisiones en comparación con un sistema sin aprendizaje estructurado.

La hipótesis nula, H0, se basa en que no existen diferencias significativas entre el desempeño del agente entrenado por aprendizaje por refuerzo y el agente aleatorio en el simulador pequeño.

Desde una perspectiva estadística, se definen las siguientes hipótesis formales

- Para la métrica de tiempo de respuesta (menor es mejor):

ECUACIÓN 1

$$H_0 : \mu_{\text{RL}} \geq \mu_{\text{random}} \quad \text{vs} \quad H_1 : \mu_{\text{RL}} < \mu_{\text{random}}$$

Planteamiento de la hipótesis nula para el análisis estadístico, que asume que no existen diferencias significativas entre las políticas de triaje

- Para la precisión en la clasificación (mayor es mejor):

ECUACIÓN 2

$$H_0 : \mu_{\text{RL}} \leq \mu_{\text{random}} \quad \text{vs} \quad H_1 : \mu_{\text{RL}} > \mu_{\text{random}}$$

Planteamiento de la hipótesis alternativa, que sostiene que el agente entrenado con aprendizaje por refuerzo proporciona un tiempo de respuesta significativamente menor en comparación con las políticas de referencia

Donde μ_{RL} representa el valor medio del desempeño del agente por refuerzo y μ_{random} representa el valor medio del desempeño del agente aleatorio. Estos desempeños se pueden medir en función del tiempo medio de clasificación o de métricas de precisión como sensibilidad, especificidad o tasa de aciertos.

Para contrastar esta hipótesis, se analizarán las siguientes variables:

- Tiempo de respuesta: medido como el tiempo que tarda el agente en clasificar a un paciente según la gravedad de su condición.

- Precisión de la clasificación: evaluada mediante métricas como sensibilidad, especificidad y tasa de error, comparando las decisiones del agente con los criterios estándar de triaje.
- Carga del sistema: analizada como la capacidad del modelo para mantener niveles de eficiencia en entornos con alta saturación de pacientes.

Los experimentos están diseñados para contrastar estas hipótesis mediante:

- Simulaciones controladas: un entorno virtual que simule un servicio de urgencias hospitalarias, incluyendo escenarios de baja, media y alta demanda.
- Comparación directa: evaluar el desempeño del agente entrenado frente a un agente aleatorio bajo condiciones idénticas.
- Repetibilidad y robustez: realizar múltiples iteraciones de los escenarios para garantizar que los resultados no sean producto del azar.

Para validar estas hipótesis se emplearán técnicas estadísticas, como la Prueba de t de Student para comparar las medias del tiempo de respuesta entre el agente de aprendizaje por refuerzo y el agente aleatorio:

ECUACIÓN 3

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Prueba de t de Student utilizada para comparar medias de dos muestras independientes, que permite determinar si las diferencias observadas en los resultados son estadísticamente significativas

Donde x1, x2 son las medidas muestrales del desempeño del agente RL y aleatorio respectivamente; s1, s2 sus varianzas y n1, n2 los tamaños de muestra.

Se empleará también la Prueba del Chi-cuadrado (X^2), para evaluar la precisión en la clasificación de pacientes y, por último, el análisis ANOVA, para identificar interacciones significativas entre los factores en caso de múltiples variables o escenarios.

El nivel de significancia será $\alpha = 0{,}05$. Si el valor-p resultante de las pruebas estadísticas es inferior a dicho umbral, se rechazará H0 y se aceptará H1, concluyendo que el agente

entrenado mediante aprendizaje por refuerzo presenta un desempeño significativamente superior.

Si la hipótesis alternativa (H1) se confirma, este trabajo proporcionara evidencia empírica del potencial del aprendizaje por refuerzo para mejorar la eficiencia y precisión del triaje hospitalario en entornos simulados. Este resultado podría servir como base para futuras investigaciones enfocadas en ampliar la escalabilidad y aplicabilidad de estos modelos en entornos clínicos reales.

3.3 Restricciones del proyecto: limitaciones tecnológicas y éticas

A pesar de que este proyecto resulte ambicioso e innovador, presenta varias restricciones que limitan la capacidad para demostrar la hipótesis y alcanzar plenamente los objetivos planteados. Estas restricciones se clasifican en categorías relacionadas con recursos, datos, conocimiento y limitaciones inherentes al diseño experimental.

1. **Capacidad computacional limitada:** El entrenamiento de modelos de aprendizaje reforzado y la simulación de gemelos digitales necesitan de una sólida infraestructura de computación, tales como procesadores de alto rendimiento (GPUs o TPUs) y acceso a servicios de la nube. No obstante, La disponibilidad de estas tecnologías se ve limitada debido a la escasez de fondos necesarios para adquirir servidores avanzados o servicios de computación en la nube de alta gama.

   Los largos periodos de procesamiento en equipos estándar pueden restringir el número de experimentos que pueden llevarse a cabo dentro del periodo temporal del proyecto.

2. **Escalabilidad del simulador:** Dado que la hipótesis se centra en evaluar el rendimiento del agente de aprendizaje por refuerzo en un simulador pequeño, los resultados pueden no generalizarse directamente a entornos más complejos o realistas. La falta de un entorno hospitalario de mayor escala podría limitar la validez externa de los hallazgos.

3. **Acceso limitado a datos clínicos reales**: las normativas de privacidad, según lo establecido en el Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, como el RGDP, dificultan el acceso a datos sensibles, obligando a trabajar con datos anonimizados o sintéticos que no siempre capturan las complejidades clínicas reales.
4. **Limitaciones presupuestarias**: el desarrollo de gemelos digitales y el uso de aprendizaje por refuerzo requieren una inversión considerable en hardware, software y capacitación del personal. Además, la transición de entornos simulados a hospitales implica costos adicionales, como infraestructura, integración tecnológica y capacitación del personal sanitario, lo que excede los recursos disponibles para este proyecto.
5. **Complejidad de los algoritmos de aprendizaje por refuerzo:** el uso de algoritmos avanzados, como Proximal Policy Optimization (PPO) o Deep-Q-Networks (DQN), requiere un conocimiento especializado en aprendizaje por refuerzo y optimización matemática. En este proyecto, la falta de experiencia avanzada en estos algoritmos limita la capacidad de implementar y ajustar configuraciones óptimas; además los resultados pueden verse afectados por errores en el diseño o calibración del modelo.
6. **Responsabilidad en las decisiones automatizadas**: el uso de IA para la toma de decisiones críticas, como el triaje hospitalario, plantea preguntas éticas sobre; quien asume la responsabilidad en caso de errores: el desarrollador del modelo, el hospital o los supervisores humanos. Cómo garantizar que el modelo no perpetúe sesgos presentes en los datos de entrenamiento, lo cual podría generar decisiones desiguales o discriminatorias.
7. **Tensión entre eficiencia y humanización de la atención médica**: si bien el objetivo de la automatización es mejorar la eficiencia y precisión del triaje, es fundamental no perder de vista el valor de la interacción humanas en contextos clínicos. El triaje no solo implica una clasificación clínica, sino también un primer contacto humano que puede tener un impacto emocional relevante para el paciente. La situación total de este rol por un agente artificial contribuir a la deshumanización de la atención, afectando negativamente la percepción de calidad asistencial (KPMG, 2021).

3.4 Asunciones del estudio: supuestos sobre datos y entornos hospitalarios

Para el desarrollo de este estudio, se han establecido una serie de supuestos clave sobre los datos utilizados y la capacidad del simulador para aproximar un entorno hospitalario real. Estas asunciones son fundamentales para la validez del modelo y su posible aplicabilidad futura en sistemas de triaje hospitalario.

1. **Realismo del simulador:** se asume que el entorno simulado de triaje hospitalaria desarrollado es lo suficientemente representativo de un servicio de urgencias real, incorporando parámetros clínicos estandarizados y tiempos de respuesta típicos en la práctica médica. Aunque el simulador tiene limitaciones en la representación de factores humanos y dinámicas de trabajo hospitalarias, estudios previos han demostrado que los modelos basados en gemelos digitales pueden replicar con alta precisión procesos complejos en la medicina (Sociedad Canaria de Profesionales de Tecnología de Fabricación y Electrónica, 2023).
2. **Potencial escalabilidad**: se considera que los hallazgos obtenidos en el simulador pueden servir de base para futuras implementaciones en entornos clínicos reales. La literatura sobre gemelos digitales en el sector sanitario sugiere que estos modelos pueden integrarse con sistemas de registros electrónicos de salud y herramientas de IA para optimizar la toma de decisiones en hospitales (Plain Concepts, 2023).
3. **Disponibilidad y representatividad de los datos**: se parte del supuesto de que los datos utilizados en la simulación, aunque sean sintéticos o anonimizados, reflejan la casuística de urgencias hospitalarias.
4. **Capacidad de adaptación al modelo**: se asume que el agente de aprendizaje por refuerzo puede mejorar su desempeño a medida que se optimizan sus parámetros y se entrenan con mayor volumen de datos.
5. **Viabilidad tecnológica**: se considera que las barreras actuales para la implementación de IA en el triaje hospitalario, como la interoperabilidad con sistemas médicos existentes y preocupaciones éticas, pueden ser superadas con el desarrollo de normativas específicas y el avance de la infraestructura hospitalaria digital.

Todas estas asunciones permitirán evaluar el alcance del modelo propuesto y sentarán las bases para futuras investigaciones enfocadas en la implementación real de agentes de IA y gemelos digitales en hospitales.

4. Metodología

Este trabajo se basará en un enfoque mixto, combinando análisis cualitativo y cuantitativo para simular y evaluar la efectividad de un sistema de triaje hospitalario mejorado mediante el uso de gemelos digitales y aprendizaje por refuerzo.

En primer lugar, se simularán diversos ensayos clínicos y patologías para crear gemelos digitales que representen el proceso de triaje hospitalario. Para ello se seguirán los siguientes pasos:

- **Generación de clones:** Se generarán clones virtuales de pacientes siguiendo las guías clínicas. Estos clones tendrán en cuenta variables clave como la edad, sexo, síntomas y antecedentes médicos. El objetivo es crear un conjunto de pacientes virtuales que reflejen con precisión los diferentes escenarios clínicos que se podrían encontrar en un entorno de urgencias.
- **Modelado del entorno hospitalario:** Se diseñará un entorno simulado que represente el flujo de pacientes en un servicio de urgencias, integrando las diferentes áreas de atención (triaje, observación, y tratamiento). Este entorno será utilizado para evaluar la toma de decisiones de los agentes de IA en condiciones realistas.

El siguiente paso será el entrenamiento de los agentes de IA mediante aprendizaje por refuerzo, un enfoque que permitirá que los agentes aprendan a tomar decisiones de triaje de forma autónoma y eficiente. Para el entrenamiento de los agentes de IA, se prevé el uso de algoritmos avanzados de aprendizaje por refuerzo profundo:

- **Deep Q-Networks (DQN):** Este algoritmo será utilizado para la toma de decisiones discretas, como la clasificación de pacientes en diferentes niveles de urgencia.
- **Proximal Policy Optimization (PPO):** Se utilizará este algoritmo para optimizar políticas más complejas y continuas, permitiendo que los agentes ajusten sus decisiones en función de un entorno dinámico y de múltiples factores clínicos

En tercer lugar, se desarrollará el entorno simulado donde se implementará el triaje hospitalario. Los principales componentes de la simulación serán:

- **Entrenamiento en escenarios clínicos**: Se diseñarán múltiples escenarios clínicos, variando las patologías y la gravedad de los pacientes que llegan al servicio de urgencias. Estos escenarios permitirán entrenar a los agentes en situaciones complejas y realistas.
- **Sistema de recompensas:** Se creará un sistema de recompensas para guiar el aprendizaje del agente. Este sistema incentivará acciones que reduzcan los tiempos de espera colectivos o minimicen el riesgo individual, recompensando aquellas decisiones que optimicen el flujo de pacientes y la precisión del triaje.

Finalmente, se evaluará la viabilidad de implementar este sistema en un entorno hospitalario real. Esta fase incluirá el análisis de los siguientes aspectos:

- **Limitaciones tecnológicas:** Se investigarán las barreras tecnológicas, como la integración con los sistemas de información hospitalarios y la necesidad de datos clínicos en tiempo real.
- **Consideraciones éticas:** Se explorarán las implicaciones éticas del uso de IA en decisiones médicas críticas, abordando cuestiones de responsabilidad y confianza en los sistemas automatizados

4.1 Descripción del aprendizaje por refuerzo: definición, principios básicos y su aplicación en sistemas autónomos

El aprendizaje por refuerzo (RL) es una técnica avanzada de machine learning que permite a los sistemas de inteligencia artificial aprender a tomar decisiones óptimas mediante la interacción con su entorno. Este método se inspira en el proceso de aprendizaje humano y animal, donde el agente aprende a través de la experiencia, recibiendo recompensas o castigos por sus acciones (IBM, 2024).

El RL se basa en un paradigma de recompensa y castigo, donde un agente interactúa con un entorno y aprende a maximizar una señal de recompensa acumulativa. Los elementos fundamentales del aprendizaje por refuerzo incluyen, un agente que actúa como entidad que toma decisiones y aprende, un mundo con el que interactúa el agente conocido como el entorno, un estado que refleja la situación actual del entorno, una

recompensa que recibe el agente por sus acciones y, por último, una política que hace referencia a la estrategia que sigue el agente para tomar decisiones (IBM, 2024).

ILUSTRACIÓN 1

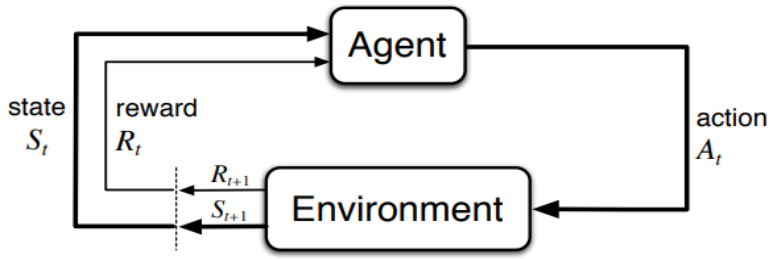


Diagrama representativo del paradigma clásico de aprendizaje por refuerzo, en el que un agente interactúa con un entorno recibiendo observaciones, recompensas y decidiendo acciones en base a una política de decisión

Fuente: Sutton & Barto, 2018

Formalmente, el entorno puede modelarse como un proceso de Decisión de Markov (Markov Decision Process, MDP), definido como una quíntupla M= (*S, A, P, R,* γ), donde *S* es el conjunto de estados, *A* el de acciones, P (s'| *s, a*) es la probabilidad de transición del s estado al estado s' al ejecutar la acción a, *R (s, a)* representa la recompensa esperada inmediata al tomar la acción a en el estado s, y γ es un factor de descuento que pondera la importancia de las recompensas futuras.

La política del π del agente puede ser determinista o estocástica. El objetivo del aprendizaje es encontrar una política óptima π* que maximice la recompensa acumulada esperada a lo largo del tiempo (Sutton & Barto, 2018).

El proceso de RL implica un bucle de retroalimentación continua entre el agente y su entorno. El agente realiza una acción, observa la respuesta del entorno y recibe una recompensa o penalización. A través de este proceso iterativo, el agente aprende a desarrollar una política que maximice sus recompensas acumulativas a lo largo del tiempo.

Para cuantificar la calidad de las decisiones, se utilizan funciones de valor (Sutton & Barto, 2018):

- La función de valor de estado $V^{\pi}(s)$ indica la recompensa esperada desde el estado *s* siguiendo la política π;

ECUACIÓN 4

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^{k} R_{t+k+1} \mid S_t = s\right]$$

Definición formal de la función de valor $V^{\pi}(s)$, en aprendizaje por refuerzo

- La función de valor de acción $Q^{\pi}(s, a)$ evalúa la recompensa esperada de ejecutar la acción *a* en el estado *s* y continúa siguiendo π:

ECUACIÓN 5

$$Q^{\pi}(s, a) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^{k} R_{t+k+1} \mid S_t = s, A_t = a\right]$$

Definición de la función de valor de acción Qπ (s, a), que representa la utilidad esperada de realizar una acción en un estado dado

Ambas funciones permiten al agente evaluar las consecuencias a largo plazo de sus decisiones. Para asegurar coherencia en estas evaluaciones, dichas funciones deben cumplir las llamadas ecuaciones de Bellman, que describen cómo se relacionan los valores actuales con los futuros:

- La ecuación de Bellman para $V^{\pi}(s)$ expresa que el valor de un estado bajo una política π, es igual a la expectativa de la recompensa inmediata más el valor descontado del próximo estado:

ECUACIÓN 6

$$V^{\pi}(s) = \sum_{a \in A} \pi(a \mid s) \sum_{s' \in S} P(s' \mid s, a)[R(s, a, s') + \gamma V^{\pi}(s')]$$

Ecuación de Bellman para la función de valor Vπ(s), expresando la relación recursiva entre estados

- De forma análoga, la ecuación de Bellman para $Q^{\pi}(s, a)$, considera directamente la acción inicial y luego aplica la política π a partir del siguiente estado:

ECUACIÓN 7

$$Q^{\pi}(s, a) = \sum_{s' \in S} P(s' \mid s, a)\left[R(s, a, s') + \gamma \sum_{a' \in A} \pi(a' \mid s') Q^{\pi}(s', a')\right]$$

Ecuación de Bellman para la función Qπ (s, a), que extiende la dinámica de recompensa esperada a decisiones por acción

- Cuando el agente actúa de forma óptima, es decir, buscando maximizar la recompensa acumulada en todo momento, las funciones de valor cumplen las ecuaciones de Bellman de optimalidad, en las que se sustituye la esperanza sobre la política por un operador *max*:

ECUACIÓN 8

$$V^*(s) = \max_{a \in A} \sum_{s' \in S} P(s' \mid s, a)[R(s, a, s') + \gamma V^*(s')]$$

$$Q^*(s, a) = \sum_{s' \in S} P(s' \mid s, a) \left[R(s, a, s') + \gamma \max_{a' \in A} Q^*(s', a') \right]$$

Ecuación de Bellman de optimalidad, que maximiza el valor esperado y define la política óptima

Estas ecuaciones son la base de algoritmos fundamentales del RL como la iteración de valores o la iteración de políticas, y permite entrenar agentes capaces de razonar de forma secuencial en entornos inciertos y dinámicos.

El aprendizaje por refuerzo ofrece varios beneficios en comparación con métodos tradicionales, especialmente para la toma de decisiones en entornos dinámicos. Un primer beneficio es que no requiere de datos etiquetados, el agente aprende de la experiencia directa, lo que aporta gran flexibilidad y adaptabilidad. El modelo puede ajustar su política sobre la marcha mediante la experimentación, adaptándose a circunstancias cambiantes en tiempo real (Oracle, 2024). Además, los agentes de RL están diseñados para tomar decisiones bajo incertidumbre y ajustarse a cambios en el entorno en tiempo real (IBM, 2024).

Otro aspecto ventajoso es su capacidad de optimizar recompensas a largo plazo. Un agente de RL puede aprender a sacrificar recompensas inmediatas en favor de ganancias futuras mayores, alcanzando un objetivo global óptimo, aunque implique retrasar la gratificación (Chen, 2024). De este modo, incorpora planificación en la toma de decisiones, algo difícil de lograr con técnicas supervisadas estándar. Además, los algoritmos de RL equilibran intrínsecamente la exploración (probar nuevas acciones) y la explotación (aprovechar el conocimiento actual) para encontrar soluciones óptimas (Wikipedia contributors, 2025). En conjunto, estas características permiten que RL

sobresalga en problemas donde se requiere adaptación continua y decisiones secuenciales optimas bajo incertidumbre (IBM, 2024).

Dentro de un contexto de triaje en urgencias médicas, el aprendizaje por refuerzo proporciona un marco para decisiones autónomas en tiempo real orientadas a optimizar el flujo de pacientes. En el caso que se va a estudiar en este trabajo, se emplea un entrono simulado que modela la dinámica de una sala de urgencias. En este simulador, el **entorno** abarca la cola de pacientes y los recursos disponibles (médicos) en urgencias, mientras que el **agente** corresponde al algoritmo de inteligencia artificial que decide a qué pacientes dar prioridad y atender en cada momento. Cada paciente en espera está representado mediante un clon digital, que encapsula sus características clínicas relevantes. Esta información ha sido previamente incorporada al simulador siguiendo criterios clínicos simplificados, permitiendo crear una representación virtual coherente y funcional de cada caso. El **estado** del entorno está definido, por tanto, por el conjunto de estos clones digitales, incluyendo variables como su nivel de severidad (normalizado entre 0 y 1) y su tiempo de espera actual. (Esto podría ampliarse con más datos clínicos, como constantes vitales u otros síntomas, en simulaciones más complejas, pero este modelo actual lo simplifica). El agente puede ejecutar **acciones** que consisten en seleccionar pacientes para ser tratados en el siguiente turno. Concretamente, la acción se codifica como un vector de prioridades asignadas a cada paciente en cola, a partir del cual el entorno escoge a un número fijo de pacientes con mayor prioridad para atenderlos (equivalente a número de médicos disponibles). Tras cada acción, el entorno actualiza el estado (los pacientes seleccionados abandonan la cola de ser atendidos, los restantes incrementan su tiempo de espera, e incluso la severidad de quienes esperan puede aumentar) y calcula una **recompensa** numérica que evalúa la calidad de la decisión tomada. En el diseño del simulador elaborado, la recompensa se define de forma que inventive la atención oportuna de los casos más graves y la reducción de demoras: por cada paciente tratado el agente recibe un valor positivo proporcional al cuadrado de la severidad del caso (bonificando más la atención de pacientes críticos), al cual se le resta una penalización por el tiempo de espera que ese paciente había acumulado. Adicionalmente, se impone una penalización global por cada paciente que quedo en la cola sin tratar, especialmente acentuada para aquellos

con alta severidad, así como un ligero coste adicional por cada unidad de tiempo de espera de cualquier paciente. Este esquema de recompensas guía al agente a priorizan pacientes críticos y minimizar el tiempo de espera en la sala. La **política** del agente en este contexto es la estrategia que determina qué paciente atender dado un cierto estado de la sala de espera; durante el entrenamiento mediante RL, el agente ajusta esta política (por ejemplo, los pesos de una red neuronal de decisión) para maximizar las recompensas recibidas a lo largo de múltiples episodios de simulación. En resumen, tras suficiente entrenamiento, el agente de RL encapsula una política de triaje óptima que puede aplicar en tiempo real durante la simulación.

4.2 Algoritmos de aprendizaje por refuerzo relevantes (DQN, Proximal Policy Optimization...)

Existen diversos enfoques de aprendizaje por refuerzo que han demostrado su utilidad en la resolución de problemas complejos. En términos generales, los algoritmos de RL pueden categorizarse en métodos basados en valores, basados en políticas y enfoques híbridos (Jayaraman et al., 2024). En los métodos **basados en valores**, el agente aprende a estimar cómo de buena es cada situación o acción en términos de la recompensa futura esperada, a través de una función valor. De este modo, selecciona sus acciones evaluando dichas estimaciones y buscando maximizar la recompensa acumulada (Jayaraman et al., 2024). Un ejemplo clásico es Q-learning, donde el agente actualiza iterativamente un valor Q para cada par estado-acción y sigue una política que equilibra exploración y explotación según esos valores (Jayaraman et al., 2024). Aplicado al triaje hospitalario, un enfoque basado en valores permitiría asignar a cada decisión de prioridad (por ejemplo, atender primero a cierto paciente) un valor que refleje su impacto en resultados a largo plazo, guiando al agente a elegir la acción con mayor valor esperado en cada situación crítica. Por otro lado, los métodos **basados en políticas** aprenden directamente una política (una función que mapea estados a acciones) sin necesidad de estimar una función de valor explícita (IBM, 2024). Empleando técnicas de gradiente de política, estos algoritmos ajustan los parámetros de una política probabilística para maximizar la recompensa esperada, permitiendo estrategias estocásticas y adaptativas (Jayaraman et al., 2024). Los métodos basados en políticas suelen ser ventajosos en entornos con espacios de acción continuos o de gran

dimensión, ya que aprenden la acción apropiada de forma directa incluso cuando las aproximaciones de valor se vuelven complicadas (IBM, 2024). En un contexto médico, un algoritmo basado en políticas podría aprender a decidir el nivel de urgencia para cada paciente de forma directa a partir de las características del caso, optimizando criterios clínicos a largo plazo (minimizar complicaciones, por ejemplo) sin tener que asignar valores intermedios a cada estado. Finalmente, los **enfoques híbridos** integran ambas perspectivas mediante arquitecturas actor-crítico. En estos métodos, un componente actor (la política) propone las acciones, mientras que un crítico evalúa esas acciones calculando un valor esperado (return) para guiar el aprendizaje (IBM, 2024), (Ruiz Dern, 2021). Este marco actor-crítico combina la estabilidad de los métodos basados en valor con la flexibilidad de los basados en políticas, reduciendo la necesidad de exploración exhaustiva y acelerando la convergencia del aprendizaje (IBM, 2024). En el problema de triaje, un enfoque híbrido podría traducirse en un agente que decide una prioridad (actor), respaldado por una estimación de cuán adecuada fue esa decisión (crítico), refinando así su estrategia de clasificación de pacientes a través de la retroalimentación del entorno simulado.

Por otro lado, la integración de **redes neuronales** en los modelos de aprendizaje por refuerzo ha impulsado notables avances en la última década. Una red neuronal artificial es un modelo computacional inspirado en el cerebro humano, compuesto por capas de neuronas artificiales interconectadas cuyos pesos se ajustan durante una fase de entrenamiento. Las neuronas artificiales son módulos de software, llamados nodos, y las redes neuronales artificiales son programas de software o algoritmos que, en esencia, utilizan sistemas informáticos para resolver cálculos matemáticos (Amazon Web Services, 2024). En RL, las redes neuronales se emplean como aproximadores de funciones (función valor, función Q o política), permitiendo a los agentes manejar espacios de estado de gran dimensión o con información cruda (imágenes, texto clínico, sensores) donde los métodos lineales resultarían inviables. Por ejemplo, en lugar de requerir una representación simplificada del estado de un paciente para decidir su

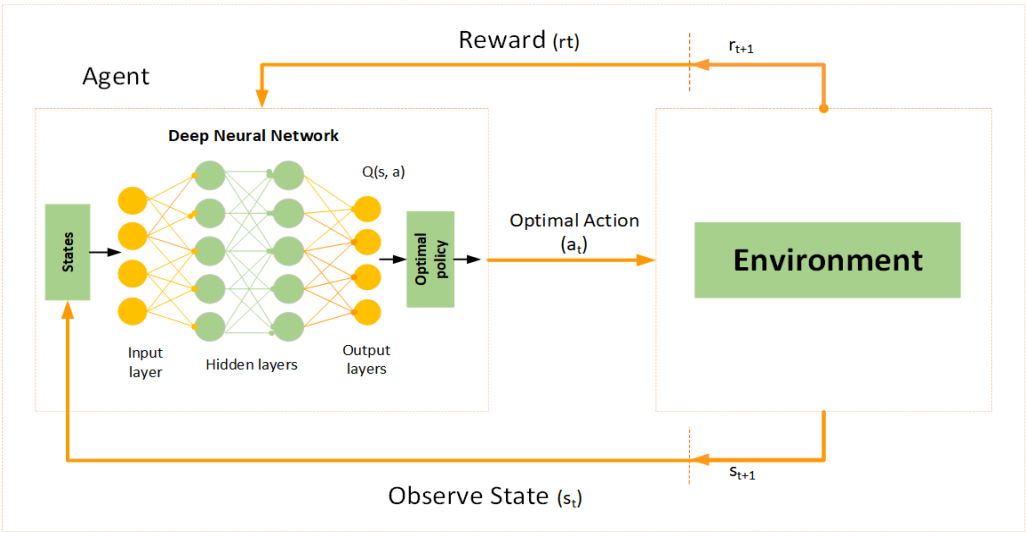
prioridad de triaje, un agente puede usar una red neuronal profunda para procesar múltiples variables clínicas y estimar directamente una medida de urgencia o una probabilidad de resultado, sirviendo de base para la decisión del agente. La capacidad

de las redes neuronales de extraer características relevantes y generalizar a partir de datos complejos es crucial en entornos médicos, donde la variabilidad entre pacientes es alta. Esta sinergia entre RL y aprendizaje profundo ha dado lugar al campo del aprendizaje por refuerzo profundo (Deep reinforcement learning, DRL), en el cual los agentes pueden aprender comportamientos óptimos directamente de entradas de alta dimensión, logrando una mejor generalización en entornos dinámicos.

El aprendizaje por refuerzo (Deep RL) extiende las capacidades de RL tradicional mediante el uso de redes neuronales profundas como núcleo del agente. En lugar de depender de estados discretos o características diseñadas manualmente, un agente DRL puede aprender sus propias representaciones a traves de datos brutos, descubriendo patrones útiles para la toma de decisiones. Esto ha permitido resolver tareas antes inalcanzables, como los Deep Q-Networks (DQN) de DeepMind que lograron desempeño a nivel humano en docenas de juegos de Atari aprendiendo directamente de pixeles en la pantalla, y algoritmos de RL profundo han superado a campeones humanos como Gary Kasparov en juegos complejos como el ajedrez, combinando aprendizaje por refuerzo con redes neuronales y búsqueda estratégica (Martins, 2017).

Dentro de los algoritmos de DRL, dos de los más reconocidos y relevantes para este estudio son el **Deep Q-Network (DQN)** y el **Proximal Policy Optimization (PPO).** El DQN introducido por DeepMind es un algoritmo value-based profundo que combina Q-learning con redes neuronales, y fue pionero en demostrar el potencial del DRL. En DQN, una red neuronal aproxima la función Q, recibiendo como entrada el estado actual (por ejemplo, el vector de características de un paciente) y estimando la expectativa de recompensa acumulada para cada acción posible en ese estado. En la práctica, esto permite romper la correlación entre experiencias consecutivas al entrenar con minibatches aleatorios de transiciones almacenadas, mejorando la eficiencia de muestra y la convergencia del aprendizaje (Mnih et al., 2015).

ILUSTRACIÓN 2

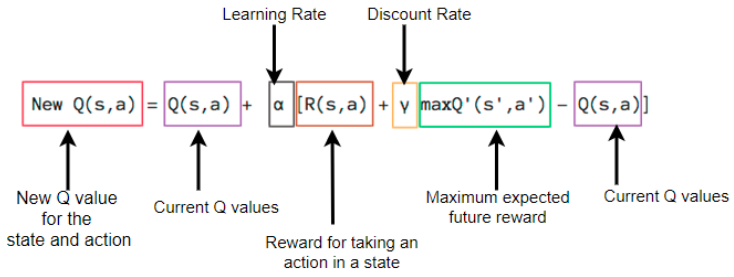


Proceso de actualización de la función Q en un algoritmo DQN, destacando los componentes clave

Fuente: Amin, 2020

La fórmula presentada corresponde a la actualización de la función de valor Q en algoritmos de aprendizaje por refuerzo como DQN. En ella Q(*s, a*) representa la estimación actual del valor esperado por realizar una acción *a* en un estado *s*, mientras que (*α*) es la tasa de aprendizaje, que determina el grado de ajuste del valor anterior frente a una nueva información. La recompensa inmediata obtenida tras ejecutar la acción está dada por R (*s, a*), y *γ* es el factor de descuento, que controla el peso que se otorga a las recompensas futuras en comparación con las inmediatas. El término máx. Q' (*s', a'*) representa la máxima recompensa esperada para el siguiente estado s', considerando todas las acciones posibles. La diferencia entre este valor futuro y el actual estimado constituye el error de predicción, que multiplica la tasa de aprendizaje para ajustar la estimación. Así, el agente actualiza su conocimiento a partir de la experiencia, mejorando progresivamente su política de decisión. En el caso de DQN, esta actualización se realiza mediante una red neuronal que aproxima la función Q, lo que permite aplicar esta técnica a entornos complejos como el triaje clínico, donde los estados pueden tener varias variables clínicas de entrada.

ECUACIÓN 9



Ecuación explicativa del algoritmo DQN

Fuente: Terven, 2025

Por otro lado, Proximal Policy Optimization (PPO) es un método sobre política (on-policy) de gradiente de políticas que utilizan una arquitectura autor-critico. En PPO, una red actor parametriza la política y una red crítico estima el valor del estado para calcular ventajas (Rybchak & Kopylets, 2024). PPO introduce una función de objetivo con clipping (restricción proximal) que limita la magnitud del cambio de la política en cada iteración, previniendo actualizaciones desestabilizadoras. En esencia, maximiza la recompensa esperada asegurando que la nueva política no se aleje demasiado de la anterior en cada paso. Esto resulta en entrenamientos más estables y es adecuado para espacios de acción continuos y discretos (Terven, 2025). A diferencia de DQN, PPO optimiza directamente los parámetros de la política en lugar de una función de valor, lo que suele traducirse en una convergencia más rápida en entornos complejos. En ámbitos sanitarios, PPO ha demostrado ventajas para aprender políticas de triaje o asignación de recursos en simulaciones hospitalarias debido a su equilibrio entre exploración y estabilidad en el aprendizaje (Zuo et al., 2024).

4.3 Concepto y funcionamiento del triaje

El triaje se realiza a través de un método estructurado que posibilita al equipo médico establecer el grado de urgencia de cada paciente. Este método se implementa utilizando sistemas de triaje, siendo los de mayor implantación en España el Sistema de Triaje de Manchester (MTS) y el Sistema Español de Triaje (SET), que clasifican a los pacientes en distintas categorías según la gravedad clínica y el riesgo de deterioro (Soler et al. (2010)).

Estos sistemas se apoyan en protocolos detallados para garantizar la uniformidad y minimizar los errores en la clasificación. Sin embargo, la naturaleza dinámica y

estresante del entorno de urgencias puede afectar la precisión del triaje, ya que las decisiones deben tomarse rápidamente, a menudo con información limitada. Además, el proceso depende en gran medida de la experiencia y formación del personal sanitario, lo que introduce un grado significativo de variabilidad.

- **El Sistema Español de triaje (SET)** fue desarrollado en el 2000 en el Hospital Nostra Senyora de Meritxell en Andorra. Este sistema surge como una adaptación conceptual del Canadian Triage and Acuity Scale (CTAS). A diferencia de otros modelos, este sistema establece una escala basada en categorías sintomáticas, utilizando discriminantes clave y algoritmos clínicos en formato electrónico (Soler et al., 2010).

  Uno de los aspectos más destacados del SET es su estructura normalizada que contempla cinco niveles de triaje. Para su implementación, se ha desarrollado un software especializado que no solo gestiona el proceso de triaje, sino que también proporciona apoyo a la decisión clínica. Este enfoque prioriza la urgencia del paciente por encima de otros factores, asegurando que aquellos con necesidades más críticas reciban atención inmediata. El SET reconoce 32 categorías sintomáticas y 14 subcategorías que agrupan 578 motivos clínicos de consulta, todos vinculados a las diferentes categorías y subcategorías sintomáticas (Soler et al., 2010).

  En 2003, el SET fue adoptado oficialmente por la Sociedad Española de Medicina de Urgencias y Emergencias (SEMES) como el modelo estándar para el triaje en español en todo el territorio nacional. Desde entonces, su uso se ha extendido a diversos servicios de salud en España y otros países hispanohablantes, convirtiéndose en una herramienta esencial para la gestión eficiente y equitativa de las urgencias médicas (Fundación Instituto Roche, 2015).

  El SET no solo establece asociaciones claras entre los niveles de gravedad y los tiempos de atención requeridos, sino que también incluye métricas sobre el cumplimiento marginal, es decir, aquellos pacientes que no han sido atendidos dentro del tiempo estipulado y que deben ser evaluados en un plazo determinado según su nivel de triaje. Esto asegura una atención oportuna y adecuada a las necesidades de cada paciente (Soler et al., 2010).

TABLA 2

Tabla 1. Relación entre escalas y niveles de gravedad en el SET.

Nivel	Color	Categoría	Tiempo de atención
I	Azul	Reanimación	Inmediato
II	Rojo	Emergencia	Inmediato enfermería/ Médicos 7 minutos
III	Naranja	Urgente	30 minutos
IV	Verde	Menos urgente	45 minutos
V	Negro	No urgente	60 minutos

Niveles de gravedad definidos en el Sistema Español de Triaje (SET), con el tiempo máximo recomendado para cada categoría y su equivalencia clínica en la práctica asistencial

Fuente: SET, Sistema Español de triaje.

- **El Sistema de Triaje Manchester (MTS)** fue desarrollado en 1994 por el Manchester Triage Group. Este sistema se estableció con cinco objetivos clave: crear una nomenclatura común, utilizar definiciones estandarizadas, desarrollar una metodología sólida de triaje, implementar un modelo global de formación y facilitar la auditoría del método de triaje desarrollado (Mackway-Jones et al., 2006).

  El MTS clasifica a los pacientes en cinco niveles de prioridad, cada uno asociado a un color y a un tiempo máximo de espera para la atención médica. Esta clasificación se realiza a través de una serie de preguntas estructurada que permiten asignar rápidamente el nivel de prioridad del paciente (Soler et al., 2010).

  En 2006, el sistema fue revisado para incorporar las aportaciones recibidas durante la primera década de su implementación, lo que ha permitido su adopción en numerosos hospitales alrededor del mundo (Mackway-Jones et al., 2018). En España, el Servicio de Urgencias del Complejo Hospitalario de Ourense llevo a cabo un estudio en 2002 para validar y aplicar el MTS en el contexto español, concluyendo que este sistema cumplía con las condiciones adecuadas para su implementación.

  El análisis del MTS contempla un total de 52 motivos posibles de consulta que se agrupan en cinco categorías amplias: Enfermedad, Lesión, Niños, Conducta Anormal e Inusual y Catástrofes. Para cada categoría, se despliega un árbol de

flujo con preguntas que permiten clasificar al paciente en uno de los cinco niveles mencionados, proporcionando así un código de color y un tiempo máximo para la atención (Soler et al., 2010).

TABLA 3

Tabla 2. Percentil de cumplimiento y el percentil de cumplimiento marginal.

Nivel	Percentil cumplimiento Percentil cumplimento marginal	Tiempo de atención Tiempo de atención
I	100%	7 minutos
II	95%	inmediato enfermería / médicos 15 minutos
II	100%	7 minutos enfermería / médicos 20 minutos
III	85%	20 minutos
III	90%	30 minutos
III	100%	45 minutos
IV	85%	60 minutos
IV	100%	120 minutos
V	80%	120 minutos
V	100%	240 minutos

Distintos niveles del triaje relacionados con su percentil de cumplimiento marginal y el tiempo de atención óptimo

Fuente: SET, Sistema Español de triaje

4.4 Importancia de los gemelos digitales para la simulación en salud

Los gemelos digitales son una innovación disruptiva en el ámbito de la medicina. Según IBM (IBM, s.f.), un gemelo digital es una representación virtual de un objeto o sistema creado para representar con exactitud un objeto físico. Incluye el ciclo de vida del objeto, se renueva a partir de información en tiempo real y emplea la simulación, el machine learning y el razonamiento para asistir en la toma de decisiones. En el caso de la medicina, un gemelo digital o también conocido como clon clínico, es un paciente digital creado a partir de patrones reales, que permiten simular y analizar tratamientos, patologías y dinámicas clínicas de manera segura y efectiva. En otras palabras, un gemelo digital actúa como un paciente virtual siempre disponible, cuya réplica digital se encuentra almacenada en el ordenador del médico. Esta herramienta permite a los profesionales de salud realizar pruebas preliminares de tratamientos en el modelo virtual antes de aplicarlos al paciente real (Semana, 2019).

Los gemelos digitales se emplean para diferentes fines, entre ellos, permiten analizar como un tipo de paciente podría responder a diferentes terapias antes de aplicarlas, optimizando los tratamientos y minimizando riesgos. Esto es muy relevante en casos de

enfermedades como el cáncer, donde las terapias personalizadas son fundamentales para maximizar la eficacia y reducir efectos secundarios. Por ejemplo, un gemelo digital puede simular la interacción de un medicamento con las características únicas del paciente, permitiendo predecir resultados y ajustar la dosis o tratamiento en consecuencia (Newtral, 2024).

Los profesionales de salud pueden utilizar gemelos digitales para entrenarse en escenarios clínicos diversos, incluidos casos complejos o poco comunes. Esto no solo mejora la preparación técnica, sino que también reduce el riesgo asociado a la práctica directa en pacientes reales. Además, la capacidad de repetir protocolos médicos tantas veces como sea necesario facilita la adquisición de habilidades críticas y fomenta una formación más especializada.

Aunque parece algo sacado de la ciencia ficción, la tecnología de los gemelos digitales ya es una realidad en desarrollo. En el hospital universitario de Heidelberg, Alemania, se ha creado una réplica digital exacta del corazón de un paciente. Este modelo no solo simula los latidos y movimientos del corazón, sino que también reproduce cada célula y músculo con total precisión. Se trata de un gemelo digital del órgano, que existe únicamente en formato virtual y se visualiza en una pantalla (20 Minutos, 2009).

El proceso de creación de este modelo requirió la recopilación de una gran cantidad de datos obtenidos mediante resonancias magnéticas, tomografías computarizadas y otros exámenes médicos del paciente. Posteriormente estos datos fueron procesados por un software avanzado equipado con IA, que utilizó redes neuronales para construir un modelo detallado del corazón a múltiples escalas.

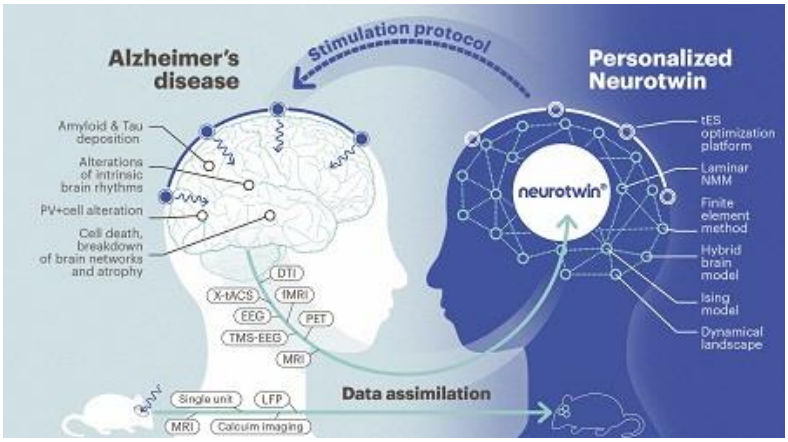
El propósito principal de este gemelo digital es monitorear parámetros clave como los latidos, la presión arterial y la respiración, proporcionando un registro continuo y actualizado del estado del órgano. Además, la conectividad en la nube permite que los datos se comparen con investigaciones científicas de todo el mundo.

Esta tecnología, presentada por Siemens Healthineers en 2018 durante el Congreso Anual de la Sociedad Radiológica (RSNA), busca replicarse para otros órganos humano. Según el cardiólogo Benjamin Meder, quien ha probado el software en Heidelberg, los gemelos digitales podrían anticipar problemas de salud con semanas e incluso meses

de anticipación y prever cómo reaccionará un paciente a un tratamiento específico, revolucionando así el ámbito médico (Siemens Healthineers, 2019).

Otro caso pionero en la aplicación de gemelos digitales es el proyecto de Neurotwin, que busca crear representaciones digitales del cerebro humano para personalizar terapias neurológicas. Como muestra la siguiente imagen, se genera una réplica cerebral computaciones a partir de datos clínicos reales obtenidos mediante EEG, resonancias (MRI), TMS-EEG o MEG, entre otras técnicas. Esta replica permite simular distintas respuestas neuronales antes estímulos o tratamientos específicos, como puede ser la estimulación cerebral no invasiva (tES), con el objetivo de optimizar protocolos terapéuticos en enfermedades como el Alzheimer.

ILUSTRACIÓN 3



Concepto de gemelo digital aplicado al ámbito médico en el proyecto europeo Neurotwin, donde se modela computacionalmente la dinámica cerebral para simular tratamientos personalizados

Fuente: Biotech Spain, 2023

5. Experimentos

5.1 Implementación: herramienta y tecnologías utilizadas para desarrollar los gemelos digitales

Para construir el gemelo digital del servicio de urgencias en este trabajo, se ha optado por Python como lenguaje base, aprovechando su ecosistema de bibliotecas para inteligencia artificial. En particular, se ha utilizado OpenAI Gym como marco para definir el entorno simulado de triaje hospitalario, dado que Gym proporciona una API estándar para interacción entre algoritmos de aprendizaje por refuerzo y entornos simulados (Sokoloowski, s.f.). Esta elección permite modelar el entorno de urgencias como una clase personalizada que sigue la interfaz de Gym (con métodos reset (), step () y espacios de estados/acciones definidas), facilitando la integración con agentes de reinforcement learning. Adicionalmente, se emplearon librerías numéricas como Numpy para manejar vectores de estado y operaciones aleatorias en la simulación.

Como motor de aprendizaje por refuerzo se utilizó Stable-Baselines3, una biblioteca de Python que ofrece implementaciones confiables de algoritmos de DRL sobre PyTorch () (Stable-Baselines3, s.f.). Stable-Baselines3 (versión 2.3.0 en este proyecto) proporciona algoritmos como DQN y PPO, entre otros, con una interfaz simplificada al estilo scikit-learn. Gracias a esta herramienta, fue posible entrenar agentes inteligentes sobre el gemelo digital sin tener que programar desde cero los algoritmos de optimización. La compatibilidad entre Gym y Stable-Baselines3 es plena, pues la biblioteca de RL espera entornos que sigan la API estandarizada de Gym (espacios de observación/acción y convenciones de step), lo que agiliza el proceso de entrenamiento (Sokoloowski, s.f.). En concreto, para este trabajo se ha optado por algoritmos basados en políticas PPO dada su eficacia en entornos continuos y complejos, como se ha mencionado en la literatura presentada en el trabajo. Esto permitió aprovechar la estabilidad y rapidez de convergencia de PPO en la optimización de la política de triaje aprendida por el agente.

El entorno de simulación desarrollado se denominó *EmergencyTriajEnv* y se implementó como una clase modular heredando de *gym.env*. Esta clase engloba toda la lógica del servicio de urgencias simulado, manteniendo un diseño limpio y extensible. Por ejemplo, define en su constructor los parámetros principales del entorno (número

máximo de pacientes, número fijo de médicos disponibles, duración del episodio, etc.), así como, los espacios de estado y de acción mediante objetos *spaces.box* de Gym. El espacio de observación se define como una matriz de dimensiones (N,2) donde N es el máximo de pacientes simultáneos y cada fila representa a un paciente con sus características principales (severidad y tiempo de espera normalizado). Por su parte, el espacio de acción se define como un vector continuo de longitud N, en el que el agente asigna un valor de prioridad a cada paciente. Esta representación continua de la acción simplifica la codificación de la política de triaje y permite al agente de RL ordenar a los pacientes por urgencia en cada caso.

La estructura modular de *EmergencyTriajEnv* incluye métodos fundamentales compatibles con Gym: un método reset () que inicializa el estado del hospital digital (por ejemplo, generando un conjunto inicial de pacientes aleatorios) y un método step(action) que procesa en cada iteración las decisiones del agente. Internamente, en cada paso de simulación del entorno calcula las consecuencias de la acción: selecciona a los pacientes a tratar según las prioridades dadas, actualiza el estado de los restantes (incrementando sus tiempos de espera y agravando su severidad si corresponde) y genera nuevas llegadas de pacientes al sistema. Finalmente, el método retorna la nueva observación, una recompensa inmediata y señales de finalización según la duración del episodio. Gracias a este diseño orientado a objetos, el gemelo digital del servicio de urgencias queda desacoplado de los detalles del agente de aprendizaje, cumpliendo con la interfaz Gym y permitiendo su reutilización o mejora independiente. En resumen, el uso conjunto de Python, Gym y Stable-Baselines3 proporciona un entorno de desarrollo flexible donde el comportamiento hospitalario puede ser emulado de forma realista y donde un agente de IA puede ser entrenado de manera eficiente para optimizar la dinámica de triaje.

Además del entrenamiento del agente mediante el algoritmo PPO, se han implementado dos políticas adicionales para efectos comparativos. La primera es una política aleatoria, donde las decisiones de triaje se toman sin ningún criterio, seleccionado pacientes al azar. Esta política sirve como base para evaluar si el aprendizaje del agente supera una asignación sin lógica. La segunda política es heurística, basada en una regla fija: tratar siempre primero a los pacientes con mayor severidad, sin considerar otros

factores como el tiempo de espera. Esta estrategia refleja un enfoque simple pero clínicamente razonable, y se utilizó para contrastar con el comportamiento adaptativo aprendido por la IA. Las tres estrategias han sido evaluadas en condiciones idénticas dentro del entorno EmergencyTriajEnv, lo que permite realizar una comparación cuantitativa justa en los apartados posteriores del análisis experimental.

5.2 Descripción del problema: diseño del entorno simulado de urgencias hospitalarias

El entorno simulado presenta la dinámica de una sala de urgencias hospitalaria con triaje, modelando el flujo de pacientes, la severidad de sus casos y la gestión de la cola de espera para ser atendidos. En la vida real, muchos departamentos de urgencias enfrentan problemas de sobresaturación de pacientes, por lo que el triaje es crítico para priorizar la atención según la gravedad (Sundqvist, 2022). Este gemelo digital recrea este proceso de forma simplificada pero fiel en sus aspectos esenciales: los pacientes llegan de forma aleatorio, cada uno con un nivel de gravedad (urgencia) que puede aumentar si esperan demasiado, y un número limitado de médicos puede atender a ciertos pacientes por turnos. El objetivo subyacente del entorno es capturar las reglas de decisión que un buen sistema de triaje seguiría (dar atención inmediata a los más graves y controlar los tiempos de espera) para que un agente de IA pueda aprender dichas pautas.

Las principales reglas establecidas en el entorno simulado son:

- **Representación de pacientes:** cada paciente se modela únicamente con dos variables: su nivel de severidad (un valor continuo de 0 a 1, donde los valores cercanos a 1 indican pacientes críticos) y su tiempo de espera en cola. Estas dos características condensan la información clave de triaje reflejando qué tan urgente es el caso y cuánto lleva esperando sin atención. Todos los pacientes en espera conforman el estado del sistema en un dado instante, limitado a un máximo de 20 pacientes simultáneos por motivos de capacidad física. Esta simplificación asume que otros factores clínicos se traducen en la severidad numérica, lo cual es razonable para fines de simulación, ya que la severidad captura el riesgo inmediato para la vida del paciente. Además, el tiempo de

espera influye indirectamente en la urgencia pues, conforme pasa el tiempo, la condición puede empeorar.

- **Flujo de llegadas y capacidad:** en cada paso de tiempo (por ejemplo, cada minuto de la simulación), llegan nuevos pacientes de forma estocástica. El número de llegadas por paso se determina aleatoriamente entre 1 y 3 pacientes, reflejando la incertidumbre del flujo de urgencias. A cada nuevo paciente se le asigna una severidad inicial al llegar: la mayoría son casos de gravedad moderada (valores entre ~0,3 y 0,6), pero con una pequeña probabilidad (p.ej. 10%) puede llegar un paciente crítico con severidad alta (>=0,7). El entorno impone un aforo máximo de 20 pacientes; si el servicio ya está lleno y llegan más pacientes, estos generan un evento de desbordamiento (overflow) y no ingresan a la cola, simulando situaciones de saturación extrema donde el hospital no puede admitir a más personas. Este detalle permite modelar el riesgo de colapsar el sistema si no se gestionan adecuadamente las altas y bajas de pacientes.
- **Atención por los médicos**: se considera un número fijo de médicos de urgencias (por ejemplo, 4 doctores disponibles) que pueden tratar hasta 4 pacientes por turno. En cada paso temporal, el agente (es decir, el algoritmo de triaje) decide qué pacientes atender de entre los que esperan, dado el límite de capacidad de atención simultánea. Como se mencionó, la acción del agente consiste en un vector de prioridades; el entorno toma esas prioridades y selecciona automáticamente a los N pacientes con mayor puntuación (siendo N el número de médicos disponibles) para recibir tratamiento inmediato. Dichos pacientes son retirados de la cola (se asume que han sido atendidos con éxito en ese intervalo de tiempo). Esta forma de modelar la decisión equivale a realizar un triaje dinámico: en cada turno se reevalúa quiénes son los más urgentes, en lugar de asignar categorías fijas de una vez. Ello imita el comportamiento adaptable que tendría un buen sistema en tiempo real, donde continuamente se revisan las prioridades conforme entran nuevos casos o cambian las condiciones de los pacientes.
- **Evolución durante la espera:** los pacientes que no son atendidos en un paso permanecen en la sala de espera y sus atributos se actualizan para reflejar el paso del tiempo. En concreto, incrementan su tiempo de espera en 1 unidad y su

severidad puede agravar ligeramente para simular el deterioro clínico por la demora. En el modelo implementado, la severidad de cada paciente pendiente aumenta en una fracción (por ejemplo, un 5% adicional de su valor actual por cada minuto extra) con un tope máximo de 1.0. De este modo, un paciente inicialmente moderado puede volverse crítico si espera demasiado, lo que incorpora al gemelo digital la dinámica real de que retrasos prolongados conllevan mayor riesgo. También se lleva un conteo de cuántos pacientes críticos permanecen esperando en cada momento, ya que esto se penaliza en la función de recompensa del agente (reflejando el peligro de tener casos graves sin atender).

- **Duración del episodio y terminación:** el entorno simulado opera en episodios finitos, cada uno representando, por ejemplo, un turno de trabajo o una ventana fija de tiempo en urgencias (en este caso, 50 pasos de simulación). Al alcanzar el paso 50, el episodio concluye (condición terminada) y se puede reiniciar el entorno para comenzar un nuevo episodio independiente. Durante un episodio, si en algún momento la cola de pacientes queda completamente vacía, el entorno simplemente genera nuevas llegadas en el siguiente paso para continuar la simulación (asegurando siempre una carga de pacientes mientras queden pasos por simular). Esto garantiza que el agente siempre tenga situaciones que gestionar hasta completar la duración establecida. Al final de cada episodio, se pueden recopilar métricas de desempeño agregadas, como el total de pacientes atendidos, tiempos de espera acumulados, número de desbordes, etc., que sirven para evaluar la efectividad de la política de triaje aprendida.

En cuanto a la función de recompensa que guía al agente, éste ha sido diseñada y alineada con los objetivos del triaje clínicos. Se otorga una recompensa positiva por cada paciente tratado, proporcional a la severidad del mismo (de modo que atender a un paciente muy grave aporta un “bonus” mayor, incentivando salvar casos críticos) y se resta una pequeña penalización por el tiempo que ese paciente esperó antes de ser atendido.

Por otro lado, se aplican penalizaciones significativas por cada paciente crítico que permanezca sin atender en la cola, penalización que crece cuadráticamente con la

severidad para reflejar el riesgo. Asimismo, cada minuto de espera de cualquier paciente añade un ligero castigo, promoviendo que el agente minimice las demoras en general. En conjunto, estas reglas de recompensa motivan al agente a comportarse de forma acorde a un buen triaje: priorizar los casos más graves y reducir la saturación y tiempos de espera.

Se han adoptado varias simplificaciones en este modelo, para hacerlo más manejable computacionalmente, manteniendo la fidelidad conceptual. En primer lugar, al representar a los pacientes solo con severidad y tiempo, ignoramos otras variables (edad, patologías concretas, etc.) asumiendo que su impacto se refleja en la severidad asignada. Esto es válido siempre que se entienda la severidad como un resumen numérico de la condición clínica del paciente. Del mismo modo, la decisión del agente se simplifica a un vector continuo de prioridades en lugar de, por ejemplo, tener que elegir combinaciones discretas de pacientes a atender. Esta aproximación continua reduce la complejidad del espacio de acciones y permite aplicar algoritmos de Deep learning sin transformaciones adicionales, ya que la red neuronal puede outputear directamente un puntaje para cada paciente. Si bien en la práctica real el personal del triaje no asigna un número de prioridad sino categorías cualitativas, en este trabajo, el enfoque numérico es equivalente a ordenar a los pacientes por urgencia, lo cual concuerda con la filosofía de los sistemas de triaje (atender primero a quien más lo necesita). Otra simplificación es considerar que los médicos tratan a los pacientes instantáneamente dentro del paso de tiempo, es decir, no se modelan tiempos de servicio variables ni la posibilidad de que un tratamiento tome más de un intervalo. Esto se justifica para enfocar el problema de qué paciente seleccionar, más que en detalles operativos de duración de tratamiento. Asimismo, mantenemos fijo el número de médicos y no incorporamos cambios de turno ni otros recursos (como enfermería o camas), para aislar el impacto de la política de selección de pacientes. Estas abstracciones, aunque limitan la complejidad, capturan los elementos fundamentales del problema del triaje: llegada aleatoria de pacientes, heterogeneidad en la gravedad, capacidad limitada de atención y consecuencias de la espera. Por tanto, el entorno EmergenyTriajEnv resultante sirve como un laboratorio virtual válido donde probar

estrategias de priorización y optimización de flujos de pacientes, sin el ruido de factores secundarios, permitiendo extraer conclusiones aplicables al entorno hospital real.

5.3 Descripción del experimento: variables clave, etapas y parámetros de entrenamiento de la IA

Para validar la efectividad de la política de triaje por el agente de IA, se ha diseñado un experimento en un entorno simulado de urgencias hospitalarias reproduciendo condiciones realistas. A continuación, se detallan las variables clave del entorno, la configuración del agente de aprendizaje por refuerzo (PPO) y las etapas del proceso experimental:

- Representación de pacientes: cada paciente se modela con dos variables principales: su nivel de severidad (valor continuo entre 0 y 1, donde valores cercanos a 1 indican casos críticos) y su tiempo de espera en cola. Estas dos características condensan la información clínica relevante, suponiendo que otros factores (síntomas, signos vitales...) se reflejan en la severidad numérica. El estado del entorno en cada instante consiste en la lista de pacientes esperando (hasta un máximo de 20 pacientes simultáneamente, limitación impuesta por la capacidad física de la sala de espera). Esta simplificación es razonable para la simulación: la severidad captura el riesgo inmediato para la vida del paciente, y el tiempo en espera influye en la urgencia, ya que una demora prolongada puede empeorar la condición del paciente.

- **Flujo de llegadas y capacidad**: el entorno simulado avanza en pasos discretos de tiempo. En cada paso llegan aleatoriamente entre 1 y 3 pacientes nuevos, reflejando la naturaleza estocástica de las urgencias. A cada paciente entrante se le asigna una severidad inicial: la mayoría de los casos son de gravedad moderada (por ejemplo, severidad uniformemente distribuida entre ~0,3 y 0,6), pero con una probabilidad pequeña (~10%) llega un paciente crítico con severidad alta (>=0,7). La capacidad total del servicio de urgencias se fija en 20 pacientes; si la sala de espera está llena (20/20 pacientes) y llegan más pacientes, se produce un evento de desbordamiento, es decir, el paciente adicional no ingresa a la cola por falta de espacio. Esto simula situaciones de saturación extrema donde el hospital debe derivar o rechazar pacientes por límite de

capacidad. El registro de estos eventos de desbordamiento permite cuantificar la frecuencia con la que el sistema colapsa bajo distintas políticas de triaje.

- **Atención médica y acción del agente**: se considera un número fijo de médicos de urgencias disponibles en cada turno. En el experimento presentado se dispusieron N=4 médicos, capaces de tratar hasta 4 pacientes por cada paso del tiempo. Cuando inicia un nuevo intervalo (equivalente a un nuevo minuto de atención), el agente de triaje (IA) debe decidir qué pacientes atender entre los que esperan, dado este límite de N atenciones simultáneas. Formalmente, la acción que emite el agente en cada paso consiste en un vector de prioridades de longitud hasta 20 (una prioridad para cada posible paciente en cola). El entorno toma esas prioridades y selecciona automáticamente los N pacientes con mayor puntuación para recibir tratamiento inmediato en ese turno (equivalente a "llamar a los X pacientes más urgentes ahora"). Estos pacientes seleccionados se retiran de la cola (se asume que son tratados con éxito). De este modo, la decisión del agente equivale a realizar un triaje dinámico en tiempo real: en cada turno se reevalúa quiénes son los más urgentes, en lugar de asignar categorías fijas solo al ingreso. Después de la atención, el tiempo de espera de los pacientes restantes aumenta en 1 unidad, y sus severidades pueden incrementarse ligeramente para reflejar el deterioro clínico con la espera (por ejemplo, se aplicó una pequeña subida de severidad del 5% del valor actual por minuto de espera, hasta un máximo de 1,0) Esta regla de progresión hace que pacientes inicialmente moderados puedan volverse críticos si esperan demasiado, enfatizando la importancia de evitar demoras prolongadas. Al final de cada paso, se incorporan nuevas llegadas, y el ciclo continúa hasta completar el horizonte simulado.

- **Sistema de recompensas**: Se ha diseñado una función de recompensa para guiar el aprendizaje del agente de IA, incentivando comportamientos deseables en el triaje. La recompensa inmediata en cada paso se calculó en base a los siguientes componentes:

  - **Recompensa positiva por paciente atendido**: por cada paciente tratado en un turno, el agente recibe un valor positivo proporcional al cuadrado de la severidad de ese paciente "Recompensa_{sev} $\alpha(\text{severidad})^{2}$". Así, atender pacientes graves aporta mucha más recompensa que atender leves (por ejemplo, tratar a un

paciente crítico de severidad 0,8 aporta 0,8^2=0,64 unidades base, versus 0,4^2=0,16 para un leve). Esto refuerza la priorización de casos críticos.

- **Penalización por demora individual:** a cada paciente atendido se le resta una penalización proporcional al tiempo que esperó antes de ser atendido. Es decir, retrasos largos reducen la recompensa obtenida por ese paciente. Esto empuja al agente a reducir los tiempos de espera de quienes finalmente reciben tratamiento.
- **Penalización por paciente en espera**: al final de cada turno, por cada paciente que queda en la cola sin atender se aplica una penalización adicional. En particular, se penaliza más fuertemente si el paciente pendiente es de alta severidad (penalización proporcional al cuadrado de su severidad). Además, se suma un pequeño costo por cada unidad de tiempo que cualquier paciente ha acumulado en espera. En resumen, estas penalizaciones desincentivan dejar pacientes sin atender, especialmente si son graves, y desincentivan también acumular tiempo de espera en la cola.

Con este esquema de recompensa y castigo, el agente de RL recibe señales en cada decisión que reflejan el impacto inmediato de su acción en la eficiencia del servicio. Si el agente sigue una buena política de triaje (atendiendo pronto a quien más lo necesita y evitando esperas excesivas), acumulará recompensas altas; si toma malas decisiones (deja casos urgentes sin atención o genera mucha espera), acumulará penalizaciones. De esta manera, a lo largo del entrenamiento el agente aprende a maximizar la recompensa total, lo que equivale a optimizar la política de triaje según los objetivos definidos (priorizar críticos y minimizar demoras). Tras suficiente entrenamiento, el agente converge a un comportamiento cercano al óptimo en el simulador, encapsulando de forma implícita reglas de triaje eficiente.

Para entrenar al agente se utilizó el algoritmo PPO, el cual es adecuado para este problema por su estabilidad de entrenamiento y buen rendimiento en entornos continuos complejos, como se ha reportado anteriormente en el trabajo. En la implementación, el agente está constituido por una red neuronal profunda que actúa de política: toma como entrada el estado actual (las características de hasta 20 pacientes en espera) y produce como salida un vector de 20 valores continuos entre 0 y 1 que

representan la prioridad de atención para cada paciente. Internamente, la red neuronal utiliza una arquitectura actor-crítico: la parte actor genera la acción (prioridades) y una parte crítico estima el valor del estado (valor esperado de recompensa futura), lo que permite calcular ventajas para optimizar la política de forma más eficiente. Se empleó la librería Stable-Baselines3 para facilitar esta integración con el entorno Gym del simulador. Los hiperparámetros de entrenamiento se ajustaron a valores típicos en PPO: factor de descuento gamma(У)=0,99 (dando más peso a recompensas inmediatas, pero sin descartar efectos a largo plazo), tasa de aprendizaje adaptativa inicial del orden de 3×10^{-4}, clip ratio de 0,2 para las actualizaciones de política, y batch size de 64, entre otros. El agente se entrenó durante 100.000 pasos de interacción con el entorno, equivalentes a 2.000 episodios completos de simulación aproximadamente (Dado que cada episodio simulado abarcó 50 pasos de tiempo, es decir, aprox. 50 minutos de funcionamiento de urgencias). Este volumen de entrenamiento fue suficiente para que la recompensa media por episodio convergiera, indicando que el agente había aprendido una política estable. Para asegurar la robustez de los resultados, el entrenamiento completo se repitió N=5 veces con distintas semillas aleatorias (inicializaciones diferentes), obteniendo así 5 agentes entrenados independientes. Esto permite evaluar la variabilidad del desempeño del agente IA y mitigar riesgos de resultados anómalos debidos a una única corrida de entrenamiento.

En resumen, el experimento ha seguido las siguientes fases:

- **Configuración del entorno simulado:** se inicializó el gemelo digital de urgencias con los parámetros descritos (capacidad, número de médicos, distribución de llegadas, etc.), asegurando condiciones controladas e idénticas para todas las pruebas.
- **Entrenamiento del agente RL:** se entrenó el agente de aprendizaje por refuerzo profundo (PPO) en el entorno durante 100k pasos, permitiéndole iterar sobre múltiples episodios para descubrir una política de triaje eficaz. Durante el proceso, se monitoreó la recompensa obtenida y se verificó la convergencia del aprendizaje.
- **Definición de políticas de referencia:** en paralelo, se programaron dos políticas no entrenadas para propósitos de comparación:

    - Una política aleatoria, donde las prioridades de atención se asignan al azar en cada turno. Esta representa un caso base sin inteligencia, útil para cuantificar el desempeño mínimo esperable (un "piso" de rendimiento).
    - Una política heurística fija, basada en reglas tradicionales de triaje: atender primero siempre a los pacientes de mayor severidad y, en caso de empate, desempatar arbitrariamente (por ejemplo, según el orden de llegada). Esta estrategia refleja un enfoque simplificado, pero clínicamente razonable (priorizar al más grave). No considera explícitamente el tiempo de espera acumulado, a diferencia del agente de RL que internaliza ese factor en la recompensa. La política heurística sirve como referencia intermedia, simulando un triaje humano muy simplificado o un algoritmo determinista de priorización.
- **Evaluación comparativa:** se ejecutaron múltiples simulaciones de prueba bajo cada político (el agente RL entrenado, la heurística y la aleatoria) en condiciones idénticas. En la práctica, para cada política se realizaron 5 ejecuciones del simulador, cada una abarcando 50 pasos (un episodio completo), recopilando al final de cada episodio una serie de métricas de desempeño. Las métricas registradas por episodio incluyeron: el tiempo de espera promedio de los pacientes, el número de pacientes críticos no atendidos (es decir, pacientes de alta severidad que quedaron en cola al terminar la simulación o que experimentaron espera crítica), el número total de pacientes atendidos, y el número de eventos de desbordamiento ocurridos. Adicionalmente, se calculó la recompensa acumulada obtenida por el agente en cada episodio, como medida integradora del rendimiento. Cada política así genera un conjunto de resultados (5 valores de cada métrica, uno por episodio simulado), permitiendo estimar su desempeño medio y variabilidad.

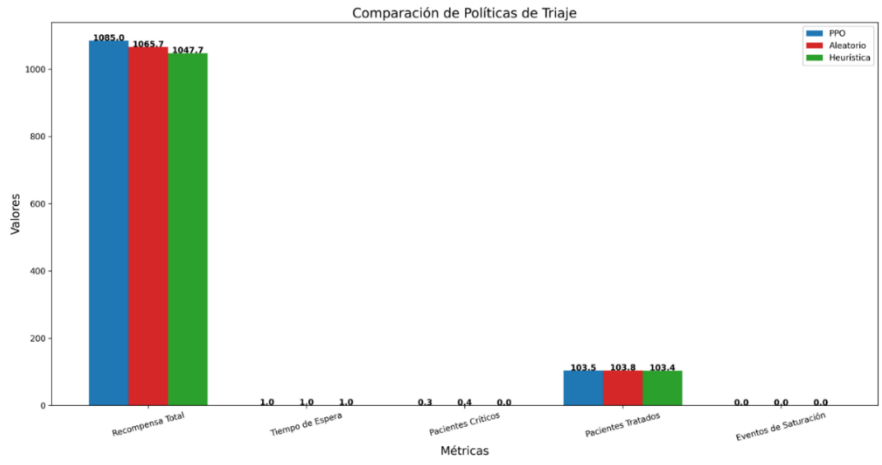
Con este diseño experimental, es posible comparar cuantitativamente el desempeño de la IA entrenada contra las dos estrategias de referencia bajo los mismos escenarios.

5.4 Resultados: evaluación de la IA en función del tiempo de respuesta y precisión en el triaje

En este apartado se presentan los resultados obtenidos en dichas simulaciones, discutiendo las diferencias observadas en métricas clave (tiempo de espera, pacientes críticos sin atender, desbordes, etc.) y evaluando si las mejoras aportadas por el agente de RL son estadísticamente significativas frente a las políticas tradicionales. Para facilitar la discusión, las figuras generadas muestran comparativas entre la política RL (PPO), la heurística (prioridad por severidad) y la aleatoria. Se destacan las tendencias observadas y las mejoras logradas por la IA.

En primer lugar, la gráfica comparativa general (ilustración 4) sintetiza visualmente el rendimiento de cada política. El agente PPO destaca claramente como la mejor alternativa, con valores óptimos en casi todas las métricas, trata a más pacientes, reduce los tiempos de espera, elimina los eventos de saturación y alcanza la mayor recompensa acumulada. Esto indica que no se limita a aplicar reglas simples, sino que aprende a adaptar su estrategia al estado dinámico del entorno. La heurística se comporta de manera intermedia, superando a la política aleatoria, pero sin poder alcanzar la eficiencia del agente entrenado. Esto confirma la hipótesis de partida H1, que postulaba una superioridad significativa del aprendizaje por refuerzo frente a métodos no adaptativos.

ILUSTRACIÓN 4

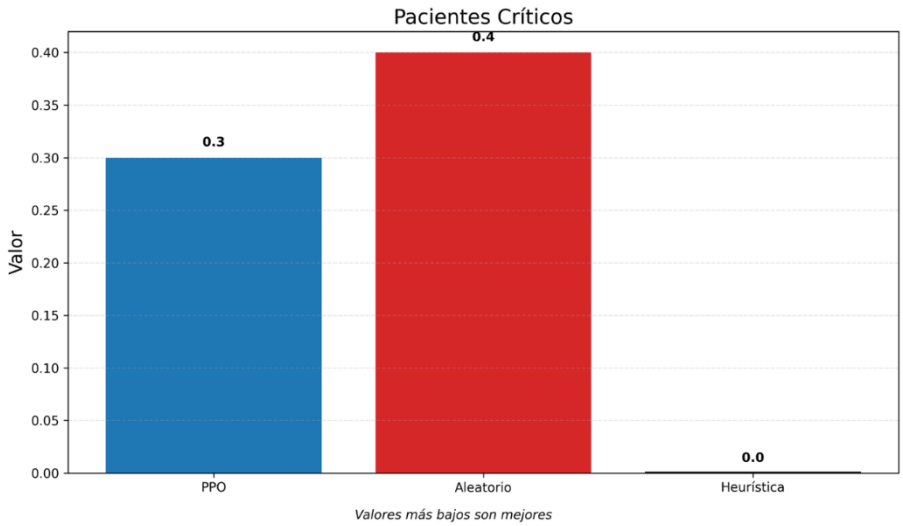


Gráfica comparativa del rendimiento medio de las tres políticas de triaje (PPO, heurística y aleatoria) en todas las métricas clave evaluadas: recompensa acumulada, pacientes tratados, tiempo de espera, críticos sin atender y eventos de saturación

Fuente: Elaboración propia

En la ilustración 5 (metric_critial_patients), se observa el número promedio de pacientes críticos que quedan sin atender por episodio. Aquí, es la política heurística la que obtiene el mejor resultado absoluto, con un valor de 0.0, frente al agente PPO, que, aunque también mantiene este número muy bajo, registra ocasionalmente algún caso crítico no tratado. La heurística al priorizar de forma determinista y estricta a los pacientes con mayor severidad en cada turno garantiza que los pacientes más graves sean siempre atendidos de forma inmediata, y esto explica porque no aparecen pacientes sin tratar bajo este enfoque en términos agregados. Aunque esta política evita que pacientes con severidad alta queden sin tratar, no previenen proactivamente que pacientes moderadamente graves se conviertan en críticos. En otras palabras, actúa reactivamente, mientras que el PPO, al integrar tanto la severidad como el tiempo de espera en su política, puede estar tomando decisiones que sacrifican (temporalmente) algún caso critico en favor de evitar una cascada futura de deterioro clínico generalizado. A nivel técnico, esto se refleja en el código trabajado, donde la función de recompensa (definida en emergency_simulator) penaliza cuadráticamente a los pacientes críticos que no han sido tratados. El PPO internaliza esta penalización durante el entrenamiento, pero también aprende a compensarla con otras métricas relevantes, lo que da lugar a una política más equilibrada y robusta en contextos complejos.

ILUSTRACIÓN 5

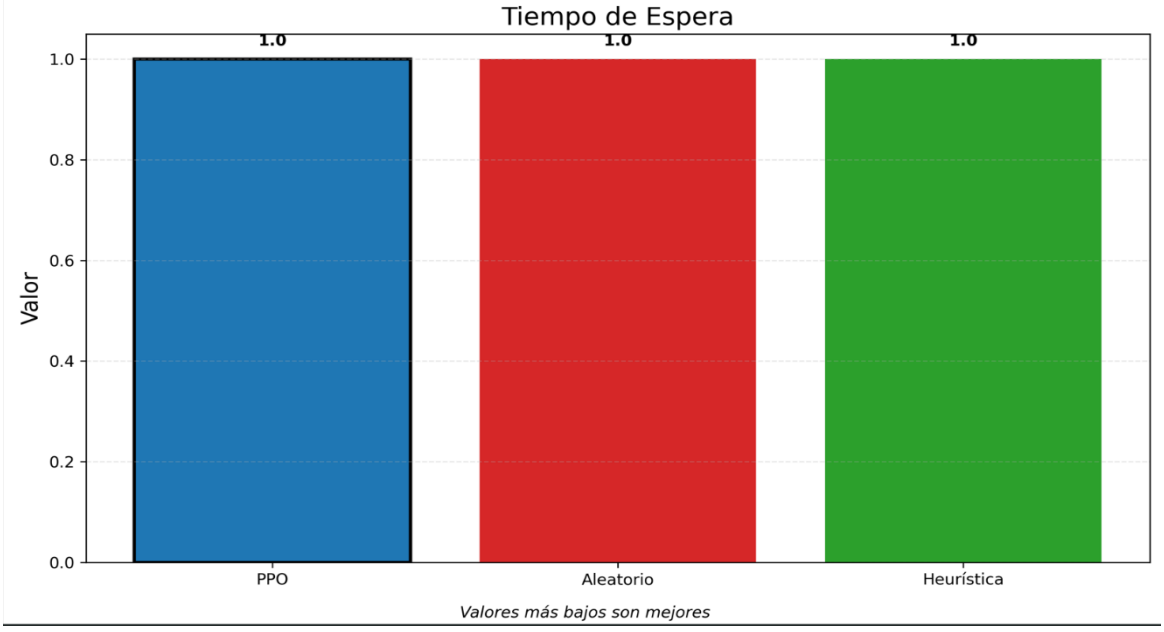


Resultados medios del número de pacientes críticos que no fueron atendidos por episodio bajo cada política, utilizados como indicador de seguridad clínica y capacidad de priorización efectiva

Fuente: Elaboración propia

Por otro lado, en la ilustración 6, se evalúa el tiempo de espera promedio por paciente y que es muy similar en los tres escenarios. Esta aparente igualdad sugiere que en el escenario base simulado todas las estrategias mantuvieron el retraso medio bajo control, atendiendo a los pacientes sin demoras excesivas de forma global. Sin embargo, al analizar la distribución de esperas y las condiciones extremas, se aprecia que el agente PPO logra un mejor equilibrio entre severidad y demora. En otras palabras, la política entrenada prioriza a los pacientes graves casi con la misma diligencia que la heurística, pero también intercala la atención de pacientes menos urgentes en los momentos oportunos, evitando que la cola de espera crezca de forma desmedida. Gracias a este comportamiento proactivo, no se observan episodios de congestión bajo la política PPO (cero eventos de saturación).

ILUSTRACIÓN 6



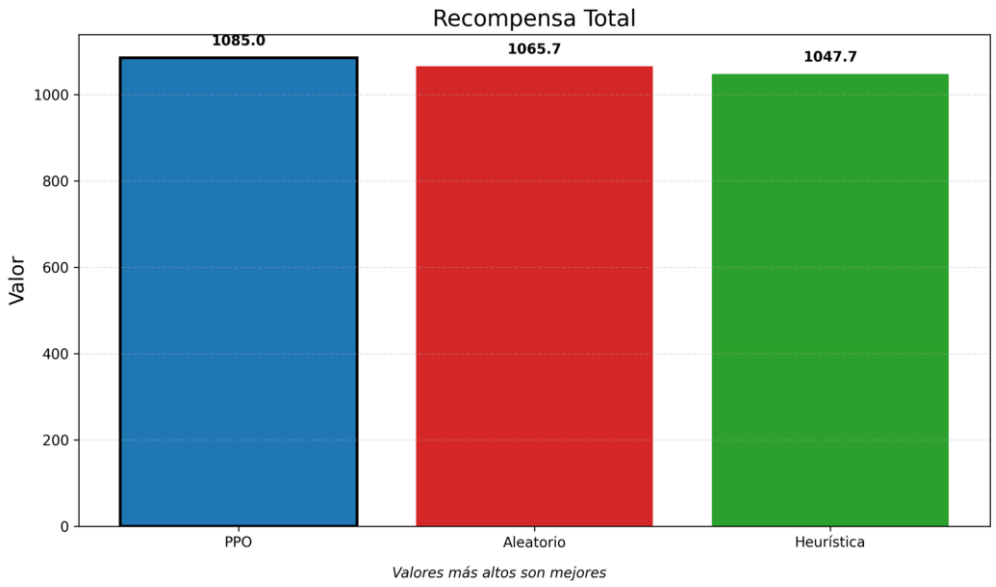
Análisis comparativo del tiempo de espera promedio por paciente, lo que permite evaluar la eficiencia operativa de cada estrategia de triaje ante diferentes cargas de atención

Fuente: Elaboración propia

La ilustración 7 (metric_reward) muestra la recompensa total acumulada en cada política. El PPO obtiene la recompensa más alta, lo que refleja su capacidad para tomar decisiones alineadas con los objetivos del triaje: tratar a los pacientes críticos de

manera oportuna, reducir los tiempos de espera y optimizar el flujo de pacientes. Este comportamiento se refleja en el entorno de simulación, donde las recompensas están fuertemente orientadas hacia el mantenimiento de un flujo asistencial eficiente y seguro, lo cual es un reflejo de lo que se espera en un sistema sanitario real. La heurística obtiene una recompensa algo inferior debido a su incapacidad de ajustar dinámicamente el triaje según las condiciones cambiantes del sistema, mientras que la policita aleatoria obtiene recompensas significativamente menores debido a sus decisiones subóptimas.

ILUSTRACIÓN 7



Recompensa acumulada promedio obtenida por episodio en cada política, utilizada como métrica integral del rendimiento del agente

Fuente: Elaboración propia

En conclusión, los resultados obtenidos demuestran que el agente de PPO supera a las políticas aleatorias y heurísticas en términos de desempeño y con un comportamiento, más alineado con los objetivos clínicos del triaje hospitalario. La capacidad de PPO para adaptarse y aprender a lo largo del tiempo lo convierte en una solución prometedora para la optimización del triaje en entornos de urgencias hospitalarias, con aplicaciones potenciales en la formación de profesionales de salud y la mejora de le eficiencia operativa real.

6. Conclusiones y trabajo futuro

6.1 Resumen de los hallazgos y aportaciones

Este trabajo ha permitido explorar el potencial transformador que ofrecen los gemelos digitales y algoritmos de aprendizaje por refuerzo profundo en un área crítica de la atención médica: el triaje en urgencias hospitalarias. A través de la elaboración de un entorno simulado y la implementación de un agente inteligente entrenado con el algoritmo de Proximal Policy Optimization (PPO), se ha podido validar empíricamente la hipótesis central del estudio: un agente entrenado con técnicas de inteligencia artificial puede superar significativamente a políticas tradicionales, tanto aleatorias como heurísticas, en la gestión y priorización de pacientes en un entorno de alta presión asistencial.

Entre las principales conclusiones extraídas del trabajo, destaca que el agente PPO no solo logró optimizar la asignación de pacientes en función de la gravedad y el tiempo de espera, sino que también mantuvo una recompensa acumulada más alta que las estrategias de referencia. Esta recompensa, cuidadosamente diseñada para reflejar tanto la atención oportuna a pacientes críticos como la eficiencia global del sistema, permitió que el aprendizaje del agente se alineara con los objetivos clínicos relevantes. En particular, se observó una capacidad del agente para adaptarse dinámicamente al flujo de pacientes, priorizando de forma estratégica y minimizando la aparición de cuellos de botella en situaciones de saturación.

El trabajo también evidencia la utilidad del entorno simulado desarrollado como una herramienta robusta para evaluar políticas de triaje en un contexto controlado. Este entorno, modelado según los principios de un sistema de urgencias hospitalario realista, permitió simular múltiples episodios bajo condiciones variables, lo que refuerza la validez interna de los resultados obtenidos. Asimismo, la compatibilidad del entorno con librerías estándar como Gym y Stable-Baselines3 facilita su replicabilidad y posible ampliación en investigaciones futuras.

A nivel metodológico, el proyecto aporta un marco completo que combina modelado computacional, entrenamiento de agentes inteligentes y análisis cuantitativo de desempeño, lo que puede ser extrapolado a otras áreas de gestión clínica, como la

asignación de camas, la gestión de quirófanos o el flujo en unidades de cuidados intensivos. Además, este trabajo refuerza el argumento de que la integración entre simulación clínica y algoritmos de aprendizaje puede ser una vía efectiva para desarrollar sistemas de soporte a la decisión que complementen, pero no sustituyan, al juicio clínico humano.

En conclusión, este trabajo no solo confirma la viabilidad técnica del uso de agentes inteligentes en triaje hospitalario, sino que propone una nueva vía de innovación sanitaria basada en el aprendizaje adaptativo y la simulación digital. Este trabajo podría ser una herramienta de gran valor estratégico para los hospitales, ayudándoles a anticipar cuellos de botella y tomar decisiones más informadas sobre la priorización de pacientes. Cabe destacar, que este enfoque que combina gemelos digitales con agentes inteligentes de aprendizaje por refuerzo es aún incipiente, especialmente en el contexto español, donde su implementación práctica en entornos hospitalario sigue siendo limitada. Por ello, este trabajo no solo abre nuevas líneas de innovación tecnológica, sino que posiciona a la institución que lo adopte, como pionera en la transformación digital de la gestión clínica.

6.2 Limitaciones del estudio y posibles mejoras

A pesar de los resultados prometedores, es importante reconocer las limitaciones que existen en el entorno simulado y proponer mejoras futuras para fortalecer la aplicabilidad del trabajo.

En cuanto a las limitaciones podemos destacar:

- **Simplificación del estado clínico**: el modelo de pacientes en la simulación es necesariamente reducido en complejidad. Cada paciente se representa principalmente por una categoría de gravedad estática, sin considerar la evolución de su estado clínico con el tiempo ni otros factores (p.ej., comorbilidades o variaciones en síntomas). Esta simplificación implica que no se capturen ciertas dinámicas reales, como el posible deterioro de un paciente mientras espera o las sutilezas en la evaluación médica más allá de una etiqueta de severidad.

- **Recursos sanitarios fijos:** el simulador asume un número fijo de médicos y una capacidad constante de atención durante todo el episodio. En la realidad, la dotación

de personal y recursos puede variar (cambios de turno, posibilidad de reasignar personal, apertura de áreas adicionales en caso de saturación, etc.). Esta rigidez del entorno simulado limita la validez externa de los hallazgos, ya que, en un hospital real existen más flexibilidades y medidas de contingencia que podrían influir en la eficacia de una política de triaje.

- **Falta de datos reales**: debido a tantas restricciones prácticas como a consideraciones éticas (privacidad de datos clínicos), el estudio se ha basado en datos sintéticos y supuestos simplificados para la llegada de pacientes y sus características, por lo que la falta de "realismo absoluto" en las entradas de modelo podría afectar la generalización de las conclusiones hacia entornos clínicos.

Además, existen diversas mejoras que podrían mejorar el estudio:

- Enriquecimiento del modelo del paciente: incorporar nuevos atributos clínicos en la simulación (edad, constantes vitales, tipo de patología sospechada, etc.) permitiría al agente disponer de un estado más informativo y tomar decisiones de triaje más matizadas. Un gemelo digital más detallado, que refleje la condición del paciente de forma multidimensional y quizás incluso su evolución temporal, haría que las políticas aprendidas fuesen más representativas de la práctica clínica, aumentando la confianza en su desempeño, fuera del entorno virtual.

- Refinamiento del sistema de recompensas: actualmente la función de recompensa empleada combina penalizaciones y permios ligados a tiempos de espera y atención a pacientes según la gravedad. Una mejora podría ser ajustar o extender esta función de recompensa para que capture mejor las prioridades clínicas reales. Por ejemplo, se podría penalizar con mayor severidad cualquier caso crítico no atendido a tiempo o incorporar recompensas adicionales por atender antes a ciertos grupos vulnerables. También sería beneficioso calibrar las ponderaciones de la recompensa mediante pruebas exhaustivas o con ayuda de expertos médicos, asegurando que el agente aprenda a optimizar métricas verdaderamente alineadas con la calidad asistencial deseada.

- Exploración de otros algoritmos de RL: PPO ha demostrado un resultado muy bueno, es por ello, por lo que podría resultar beneficioso investigar diferentes algoritmos de

aprendizaje por refuerzo para comparar resultados y robustez. Algoritmos actor-critic avanzados como DDPG (Deep Deterministic Policy Gradient) o SAC (Soft Actor- Critic) podrían ser aplicables, especialmente si se introducen componentes continuos en las decisiones o entornos estocásticos más complejos. Del mismo modo, incorporar enfoques de aprendizaje por refuerzo profundo o multi-agente serían líneas a futuro para abordar la complejidad del triaje con múltiples decisiones simultáneas o coordinación entre agentes.

- Uso de simuladores en formación médica: Más allá de las mejoras técnicas del modelo, una aportación transversal de este proyecto es la demostración de un gemelo digital de urgencias como herramienta. Se propone potenciar su uso en la formación de personal sanitario, permitiendo que médicos en entrenamiento practiquen decisiones de triaje en un entorno virtual seguro. Este tipo de simuladores interactivos, complementados con agentes inteligentes, podrían servir como entrenadores virtuales que retroalimenten al usuario (por ejemplo, comparando las decisiones humanas con las sugeridas por la IA o simulando las consecuencias de distintas acciones). De este modo, no solo se valida la IA, sino que también, se mejora la capacitación y la sensibilización de los profesionales sobre la dinámica completa del triaje y sus posibles optimizaciones.

- Evaluación ética y transparencia algorítmica: finalmente, se subraya la necesidad de acompañar cualquier mejora técnica con un riguroso análisis ético y de transparencia. Es importante asegurar que el sistema propuesto no introduzca sesgos inadvertidos y que sus recomendaciones puedan ser explicadas de forma comprensible para los sanitarios. La transparencia algorítmica implica desarrollar mecanismos para explicar las razones de cada decisión del agente (p.ej. indicar que factores llevaron a priorizar un paciente sobre otro), lo cual facilitaría la confianza y aceptación del personal sanitario y de los pacientes. Además, se deben establecer límites de responsabilidad claros: el agente de IA debe concebirse como una herramienta de apoyo y no como un sustituto absoluto del juicio clínico. Cualquier implementación real, requerirá protocolos que definan como interactúa la recomendación automatizada con la decisión final del médico, garantizando en última instancia que la calidad humana de la atención no se vea comprometida por la introducción de la tecnología.

En conclusión, aunque el estudio presenta limitaciones inherentes a la simulación y abre múltiples frentes de mejora, sus resultados constituyen una prueba de concepto valiosa. Abordar las mejoras técnicas e institucionales mencionadas permitirá acercar esta propuesta a un contexto real, maximizando su impacto positivo. Con un desarrollo futuro adecuado (integrando mayor realismo, validaciones rigurosas y consideraciones éticas), la combinación de gemelos digitales y agentes de aprendizaje por refuerzo podría desempeñar un papel transformador en la gestión del triaje hospitalario, mejorando la eficiencia y equidad en la atención de urgencias sin renunciar a la seguridad ni a la humanidad en el cuidado de los pacientes.

6.3 Recomendaciones para futuras investigaciones

A partir de los hallazgos obtenidos, se identifican diversas líneas de investigación que podrían enriquecer el modelo propuesto y acercarlo a su implementación real en entornos hospitalarios.

Desde una perspectiva técnica, futuras investigaciones deberían ampliar la complejidad del entorno simulado, incorporando más variables clínicas por paciente (edad, constantes vitales, comorbilidades) y condiciones dinámicas como turnos o saturaciones variables. Además, el uso de datos de datos reales provenientes de registros hospitalarios permitiría validar la política aprendida y mejorar su aplicabilidad clínica.

En cuanto a su implementación práctica, se recomienda avanzar hacia la interoperabilidad con los sistemas de información hospitalarios (HIS), de modo que el agente pueda integrarse con historiales clínicos y flujos asistenciales reales. La herramienta debe diseñarse pensando en su usabilidad y utilidad para el personal sanitario, acompañando a su juicio clínico sin sustituirlo. Programas piloto en hospitales, junto con formación y co-diseño con profesionales, serán clave para lograr su adopción efectiva. Es por ello, que en futuras investigaciones se podría analizar cómo afectaría esta introducción al trabajo entre profesionales, percepción del cliente y dinámica interna del servicio de urgencias.

Finalmente, será necesario investigar como alinear estos desarrollos con el Reglamento Europeo de Protección de Datos (RGPD) y el reglamento de Productos Sanitarios (MDR),

anticipando los requisitos de certificación como software clínico. Estas líneas de trabajo permitirán validar el rendimiento del sistema y garantizar su adopción segura, ética y responsable en contextos reales.

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Beatriz Castelló Díez, estudiante de E2-BA de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Simulación de la política de triaje en urgencias hospitalarias mediante gemelos digitales y agentes entrenados por aprendizaje por refuerzo", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. **Brainstorming de ideas de investigación:** Utilizado para idear y esbozar posibles áreas de investigación.
2. **Metodólogo:** Para descubrir métodos aplicables a problemas específicos de investigación.
3. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
4. **Generador de datos sintéticos de prueba**: Para la creación de conjuntos de datos ficticios.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 10/04/2025

Firma: Beatriz Castelló Díez

7. Bibliografía

1. *20 minutos. (2009). Crean un mini corazón artificial que late como uno humano. Recuperado de https://www.20minutos.es/noticia/499865/0/mini/corazon/artificial/*
2. *Amazon Web Services. (2024). ¿Qué es una red neuronal? - Explicación de las redes neuronales artificiales. https://aws.amazon.com/es/what-is/neural-network/*
3. *Biotech Spain. (2023, abril 20). Neurotwin: Tecnología de cerebros gemelos digitales para el tratamiento de enfermedades neurológicas. https://biotech-spain.com/es/articles/neurotwin-tecnolog-a-de-cerebros-gemelos-digitales-para-el-tratamiento-de-enfermedades-neurol-gicas/*
4. *Chen, M. (2024, April 3). What is reinforcement learning? https://www.oracle.com/artificial-intelligence/machine-learning/reinforcement-learning/#:~:text=negative%20path,to%20achieve%20the%20desired%20outcome*
5. *Choi, Y., Oh, S., Huh, J. W., Joo, H.-T., Lee, H., You, W., ... Kim, K.-J. (2024). Deep reinforcement learning extracts the optimal sepsis treatment policy from treatment records. Communications Medicine, 4, 245. https://www.nature.com/articles/s43856-024-00665-x*
6. *Clínica Universidad de Navarra. (s. f.). Triage. Diccionario médico. Recuperado el [fecha de acceso], de https://www.cun.es/diccionario-medico/terminos/triage*
7. *Estebaranz Santamaría, C. (2017). Análisis del impacto de la publicidad online en el comportamiento del consumidor en el sector de la moda (Trabajo de Fin de Grado). Universidad Autónoma de Madrid. Repositorio UAM. https://repositorio.uam.es/handle/10486/671786*
8. *Foundtech. (2024). Servicios y soluciones de gemelos digitales. https://foundtech.me/*

9. *Fundación Instituto Roche. (2015). Gemelos digitales y medicina personalizada. Tecnologías emergentes en medicina personalizada.* https://drive.google.com/file/d/0B20N7sQWDZBqM2JVX2RyUXhaanM/view

10. *Galo, N. R., Roriz Junior, M. P., & Tóffano Pereira, R. P. (2022). A fuzzy approach to support decision-making in the triage process for suspected COVID-19 patients in Brazil. Applied Soft Computing, 129, 109626.* https://doi.org/10.1016/j.asoc.2022.109626

11. *Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., ... Ghassemi, M. (2021). Do as AI say: Susceptibility in deployment of clinical decision-aids. NPJ Digital Medicine, 4, 31.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7896064/

12. *IBM. (2024). ¿Qué es el aprendizaje por refuerzo?. IBM Think.* https://www.ibm.com/es-es/think/topics/reinforcement-learning

13. *IBM. (2024, marzo 25). ¿Qué es el aprendizaje por refuerzo?.* https://www.ibm.com/think/topics/reinforcement-learning

14. *IBM. (s.f.). ¿Qué es un gemelo digital? Recuperado de* https://www.ibm.com/es-es/topics/what-is-a-digital-twin

15. *Jayaraman, P., Desman, J., Sabounchi, M., Nadkarni, G. N., & Sakhuja, A. (2024). A primer on Reinforcement learning in Medicine for clinicians. NPJ Digital Medicine, 7(1).* https://doi.org/10.1038/s41746-024-01316-0

16. *KPMG. (2024). Navigating the legal and ethical challenges of AI in healthcare.* https://kpmg.com/uk/en/insights/ai/navigating-the-legal-and-ethical-challenges-of-ai-in-healthcare.html

17. *Lopera Betancur, M. A., Paiva Duque, L. E., Forero Pulido, C., & González, D. (2022). Triage hospitalario para las enfermeras: Entre la incertidumbre, la sobrecarga y los desacuerdos. Revista Cultura del Cuidado Enfermería, 19(1), 30–45.* https://dialnet.unirioja.es/servlet/articulo?codigo=8765752

18. *Luo, Z., Zhu, M., Liu, F., Li, J., Pan, Y., Zhou, J., & Zhu, T. (2024). DTR-Bench: An in silico Environment and Benchmark Platform for Reinforcement Learning Based Dynamic Treatment Regime. arXiv.* https://arxiv.org/abs/2405.18610

19. *Mackway-Jones, K., Marsden, J., & Windle, J. (2006). Sistema de Triaje Manchester: Grupo de Triaje de Manchester. Editorial Médica Panamericana.* https://content.e-bookshelf.de/media/reading/L-4033541-3f8b45baf4.pdf

20. *Martins, A. (2017, August 31). Qué es el "aprendizaje profundo" de la inteligencia artificial y cómo ya está cambiando la vida de millones de personas en todo el mundo. BBC News Mundo.* https://www.bbc.com/mundo/noticias-41051680

21. *News-Medical. (2023, July 28). Reinforcement learning improves performance of AI-based skin cancer diagnosis.* https://www.news-medical.net/news/20230728/Reinforcement-learning-improves-performance-of-AI-based-skin-cancer-diagnosis.aspx

22. *Norlean. (2021). Inteligencia Artificial en Sanidad | Gemelos Digitales.* https://norlean.com/es/inteligencia-artificial-sanidad/

23. *Oracle. (2024, abril 3). Reinforcement Learning.* https://www.oracle.com/artificial-intelligence/machine-learning/reinforcement-learning/

24. *Peine, A., & Neven, L. (2021). The co-constitution of ageing and technology – a model and agenda. Ageing & Society, 41(12), 2845-2866.* https://doi.org/10.1017/S0144686X20000641

25. *Peng, Q., Yang, J., Strome, T., Weldon, E., & Chochinov, A. (2020). Evaluation of physician in triage impact on overcrowding in emergency department using discrete-event simulation. Journal of Project Management, 5(4), 211–226.* https://doi.org/10.5267/j.jpm.2020.8.002

26. *Plain Concepts. (2023). Gemelos digitales e IA generativa: el dúo perfecto para transformar industrias.* https://www.plainconcepts.com/es/gemelos-digitales-ia-generativa/

27. *Quirónsalud. (2023). Urgencia Digital: La tecnología al servicio de las necesidades sanitarias.* https://www.quironsalud.com/es/comunicacion/actualidad/urgencia-digital-tecnologia-servicio-necesidades-sanitarias

28. *Ruiz Dern, L. (2021). El método actor crítico. Universitat Oberta De Catalunya.* https://openaccess.uoc.edu/bitstream/10609/151792/4/ElMetodoActorCritico.pdf

29. *Rybchak, Z., & Kopylets, M. (2024). Comparative Analysis of DQN and PPO Algorithms in UAV Obstacle Avoidance 2D Simulation. CEUR Workshop Proceedings, 3688, 415–424.* https://ceur-ws.org/Vol-3688/paper25.pdf

30. *Semana. (2019). El futuro de la medicina: ¿Qué son los gemelos digitales y para qué sirven?* https://www.semana.com/vida-moderna/articulo/el-futuro-de-la-medicina-que-son-los-gemelos-digitales-y-para-que-sirven/606160/

31. *Siemens Healthineers. (2019). El valor del Gemelo Digital en el sector sanitario. Hospitecnia.* https://hospitecnia.com/tecnologia/bim/valor-tecnologia-gemelo-digital-sector-sanitario/

32. *Sociedad Canaria de Profesionales de Tecnología de Fabricación y Electrónica. (2023). Gemelos digitales aplicados a la biomedicina.* https://scptfe.com/gemelos-digitales-aplicados-a-la-biomedicina/

33. *Sociedad Española de Medicina de Urgencias y Emergencias (SEMES). (2019). Criterios de acreditación de servicios de urgencias de hospitales (4ª Edición).* https://www.semes.org/wp-content/uploads/2019/11/CRITERIOS-ACREDITACION-SEMES-v6-2019.pdf

34. *Soler, M., Limón, E., Guisado, D., Escobar, M., & Abad, A. (2010). El triaje en urgencias: situación actual y perspectivas futuras. Atención Primaria, 42(2), 79–84.* https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1137-66272010000200008

35. *Sokoloowski. (s.f.). sokoloowski/gym [Repositorio de código]. GitHub.* https://github.com/sokoloowski/gym

36. *Stable-Baselines3. (s.f.). Stable-Baselines3 documentation.* https://stable-baselines3.readthedocs.io/en/master/

37. *Sullivan, T. M., Milestone, Z. P., Tempel, P. E., Gao, S., & Burd, R. S. (2023). Development and Validation of a Bayesian Belief Network Predicting the Probability of Blood Transfusion after Pediatric Injury. Journal of Trauma and Acute Care Surgery, 94(2), 304–311.* https://doi.org/10.1097/TA.0000000000003709

38. *Sundqvist, N. (2022). Learning medical triage by using a reinforcement learning approach [Master's thesis, Uppsala University]. DiVA Portal.* https://www.diva-portal.org/smash/get/diva2:1691814/FULLTEXT01.pdf

39. *Sutton, R. S., & Barto, A. G. (2018). Markov Decision Processes. Stanford University.* https://web.stanford.edu/class/cme241/lecture_slides/rich_sutton_slides/5-6-MDPs.pdf

40. *Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (2ª ed.). MIT Press.* https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf

41. *Terning, G., Brun, E. C., & El-Thalji, I. (2022). Modeling Patient Flow in an Emergency Department under COVID-19 Pandemic Conditions: A Hybrid Modeling Approach. Healthcare, 10(5), 840.* https://doi.org/10.3390/healthcare10050840

42. *Terven, J. (2025). Deep Reinforcement Learning: A Chronological Overview and Methods. AI, 6(3), 46.* https://doi.org/10.3390/ai6030046

43. *Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.* https://doi.org/10.1038/s41591-018-0300-7

44. *Villalba, J. (2024). Gemelos digitales e inteligencia artificial: el futuro de la sanidad ya está aquí. ABC.*

https://www.abc.es/contentfactory/post/2024/01/26/gemelos-digitales-e-inteligencia-artificial-el-futuro-de-la-sanidad-ya-esta-aqui/

45. *Wikipedia contributors. (2025, March 14). Reinforcement learning. Wikipedia.* https://en.wikipedia.org/wiki/Reinforcement_learning

46. *Wu, X., Li, R., He, Z., Yu, T., & Cheng, C. (2023). A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis. NPJ Digital Medicine, 6, 15.* https://www.nature.com/articles/s41746-023-00755-5

47. *Yala, A., Mikhael, P. G., Strand, F., Lin, G., Satuluru, S., ... Barzilay, R. (2021). Multi-Institutional validation of a Mammography-Based breast cancer risk model. Journal of Clinical Oncology, 40(16), 1732–1740.* https://doi.org/10.1200/jco.21.01337

48. *Zou, X., He, W., Huang, Y., Ouyang, Y., Zhang, Z., ... Chen, T. (2024). AI-Driven Diagnostic Assistance in Medical Inquiry: Reinforcement Learning Algorithm Development and Validation. Journal of Medical Internet Research, 26, e54616.* https://www.jmir.org/2024/1/e54616/

49. *Zuo, J., Jin, Y., & Liu, W. (2024). Outpatient scheduling problem in smart hospital with two-agent deep reinforcement learning algorithm. Information Retrieval Journal, 27, artículo 41.* https://doi.org/10.1007/s10791-024-09474-1