Detección de anomalías en pacientes hospitalarios con el dataset MIMIC III

Isaac Esteban Uribe Jaramillo

Ingeniería, Universidad de Antioquia, Medellín, Antioquia, Colombia;

Correo: isaac.uribej@udea.edu.co

Keywords: vital signs prediction, anomaly detection, patient monitoring, early warnings

Abstract

Proporcionar una intervención temprana y un trato humano debe ser una prioridad en el servicio médico. Por eso, ofrecer métodos y herramientas analíticas que apoyen al personal sanitario es clave para mejorar la atención en hospitales y UCI. Este estudio usó la base de datos MIMIC-III para analizar signos vitales clave: frecuencia cardíaca, saturación de oxígeno, temperatura y presión arterial. Se filtró y preprocesó un conjunto grande de datos, organizándolo cronológicamente por paciente y muestra. Se creó una etiqueta artificial “PATIENT_STATE” para clasificar los registros en normal, taquicardia o bradicardia. Con técnicas de limpieza e imputación KNN, el dataset final tuvo 72,922 registros.

El análisis mostró distribuciones similares entre grupos y una fuerte correlación esperada entre presiones sistólica y diastólica. La detección con Factor Local Outlier identificó anomalías clínicas realistas en glucosa y temperatura, captando detalles que otros métodos simples no detectaban. Aun hace falta mejorar tanto calidad y análisis de datos, este enfoque promete alertar oportunamente sobre condiciones críticas, que podria brinda una monitorización más inteligente y automatizada, lo cual, puede reducir costos hospitalarios al prevenir complicaciones durante los ingresos.

1. Introducción

En un contexto clínico la detección de mediciones anómalas en signos vitales o análisis clínicos es de un interés valioso en el área médica, pues con la detección temprana de signos vitales que superen los rangos normales, se pude generar una alerta para informar a los profesionales de la salud sobre posibles deterioros en el estado del paciente, pues si diéramos tratamiento pacientes sin determinar sus signos vitales, es posible que no reflejemos la urgencia de su situación [1]. El grado de anomalías y las patologías presentes en un paciente cuando es ingresado también nos puede suministrar información de cual debería ser el comportamiento de sus demás signos vitales, como la evolución de este.

El combinar métodos de análisis y predicción resulta importante pues esto permite el manejo de recursos de una manera más efectiva, la atención temprana y un trato mas humano por parte del personal médico. Además, las ventajas que ofrecen las implementaciones de monitoreo de signos vitales con modelos predictivos, posibilitan las intervenciones oportunas, que pueden tener un alto impacto en el diagnóstico y la prevención de enfermedades y complicaciones durante los ingresos hospitalarios [2]. Esto también reduce los costos hospitalarios al disminuir la necesidad de tratamientos innecesarios y ayudar a remediar problemas en futuros reingresos. De esta manera, estos modelos permitiran personalizar el cuidado según las necesidades específicas de cada paciente.

Según el protocolo de monitoreo no invasivo de signos vitales de la cínica sagrado corazón, "*La temperatura, la respiración, la frecuencia cardiaca, la saturación, la tensión arterial y la frecuencia cardiaca fetal, son parámetros a través de los cuales es posible evaluar el estado hemodinamico de la salud de un individuo, pues sus valores se mantienen constantes dentro de ciertos límites, en estado de normalidad*" [3], evidenciando que mantener un monitoreo constante y preciso de estos parámetros es fundamental para detectar cualquier desviación que pueda indicar un deterioro en la salud del paciente, además aportando una visión objetiva de cuales marcas se deberían tener presentes para el análisis oportuno del estado de un paciente, siendo su temperatura corporal, su frecuencia cardiaca y porcentaje de oxigeno en la sangre características fundamentales en la predicción del estado y la predicción de este; por lo tanto, la medición precisa de estos signos requiere práctica y tiempo, lo que presenta una limitación importante en los métodos tradicionales de monitoreo clínico. Además, el monitoreo constante de todos los pacientes que es una tarea casi imposible por la limitación del personal medico disponible para la atención de muchos pacientes.

Estas restricciones introducen la necesidad de desarrollar nuevas técnicas y tecnologías que permitan el monitoreo automatizado, la predicción y el apoyo en la evaluación del estado de salud de un paciente. La incorporación del análisis de datos ayudará a solventar las problemáticas derivadas de los métodos convencionales y proporcionará soporte en el análisis de datos y tendencias médicas. Mediante técnicas de imputación, limpieza y tratamiento de datos de pacientes reales, se intentará sistematizar la detección de datos anómalos en pacientes con indicios o diagnóstico de taquicardia o bradicardia, conceptos definidos en el protocolo de monitoreo [3][4]

como trastornos o alteraciones en el ritmo cardíaco por encima de 100 bpm y por debajo de 60 bpm, respectivamente. Con estas etiquetas se busca llevar a cabo labores de análisis clínico para entender cómo se correlacionan los signos vitales y cómo estos explican comportamientos anómalos en pacientes con determinadas condiciones.

2. Metodología

En primera instancia se adquirió la base de datos MIMIC-III para el desarrollo de esta actividad. Se revisaron los archivos que forman parte del compendio suministrado junto con el dataset, en particular el archivo CHARTEVENTS.csv, que contiene varias columnas, entre ellas un identificador único para cada dato tomado a un paciente. Este identificador está asociado a un código que especifica qué signo vital fue medido en cada registro.

En base a esta información se realizó un filtrado y una preselección de los signos vitales relevantes para el análisis, decidiendo trabajar únicamente con la información contenida en el archivo CHARTEVENTS.csv y se acoto a seleccionar signos vitales. Esto se hizo para evitar la complejidad de manejar y concatenar información dispersa en varios archivos, lo cual complicaría el pre-procesamiento y procesamiento de datos.

Proceso de acotamiento y preparación del dataset:

Debido al gran tamaño del dataset original de MIMIC-III, fue necesario realizar un proceso de acotamiento y preparación para hacer los datos manejables para este análisis. El archivo contiene una columna con los valores de las mediciones de los biomarcadores y otra con el identificador o código que especifica el tipo de biomarcador (Ritmo cardíaco, saturación de oxígeno, yemperatura corporal, glucosa, presión sistólica y presión diastólica).

Para trabajar con las variables de forma individual, se siguió el siguiente procedimiento:

- **Procesamiento por biomarcador:** Se procesó el dataset original en fragmentos (chunks) para controlar el volumen y evitar sobrecarga de memoria.
- **Filtrado por código:** En cada fragmento se seleccionaron únicamente los biomarcadores de interés, identificados por sus códigos únicos.
- **Extracción y almacenamiento:** Los datos filtrados para cada biomarcador se extrajeron y guardaron en archivos CSV separados.
- **Ensamblaje del dataset final:** Posteriormente, los datos extraídos de cada biomarcador fueron combinados en un único DataFrame.
- **Organización cronológica:** Finalmente, el dataset ensamblado se ordenó por la fecha y hora de toma de muestra (CHARTTIME), para asegurar un análisis temporal preciso de la

evolución de los biomarcadores, ademas se ordeno por paciente los cuales tienen un “id” unico asociado a ellos.

El dataset original contaba con 9028427 entradas, pero se decidio reducirlo a 100.000 registros para que el procesamiento no fuera tan pesado y no llevara mucho tiempo a la hora corer los notebooks. Posteriormente con el fin de acotar el problema, se creo una variable objetivo o variable de observabilidad, para dar mas explicabilidad a los datos y haremos uso de una función para crear esta columna con una "etiqueta artificial" (PATIENT_STATE) que representa el estado clínico estimado del paciente durante la remisión o ingreso hospitalario. Esta etiqueta busca reflejar si el ingreso fue normal o si el paciente parecía presentar o si se sospechaba que podía tener alguna condición crítica según sus signos vitales.

La función asigna el estado clínico a cada registro en base a umbrales clínicos comúnmente aceptados.[3] de esta manera pues se clasifico la población en 3 clases pacientes con: “Taquicardia”, “Braquicardia” y “**Normal**”, con los cuales se procedio hacer un analisis exploratorio con el cual se empezo por consultar el disbalance de las clases, de los cuales se podia observer que el 63.27% de las mediciones de pacientes presentaba un estado Normal en su ritmo cardiaco durante su ingreso hospitalario, un 33.84% de las mediciones correspondian a un aumento del ritmo cardiaco y solo un 2.89% eran registros asociados a una disminucion del mismo.

Se hizo primero una normalizacion en los datos de temperature que presentaban valores con mediciones extremas con temperaturas superiors a 100°, se concluyo que probablemente estuvieran en grados Farenheit que es una unidad de medicion comun en EEUU y se decidio transformarlo usando la formula (1)

$$°C = (°F - 32) \times \frac{5}{9} \quad (1)$$

Después se realizó un análisis de la distribución de los datos para cada una de las categorías acotadas anteriormente, efectuando un análisis univariado para cada variable involucrada en el problema. Se encontraron distribuciones similares con ligeros cambios entre las categorías. Adicionalmente, el análisis multivariado exploró la correlación entre variables, encontrándose una correlación fuerte únicamente entre la presión sistólica y la presión diastólica, lo cual es esperable dado que ambas representan medidas relacionadas del ciclo cardíaco: la presión sistólica corresponde a la fuerza con la que el corazón bombea la sangre, mientras que la diastólica es la presión cuando el corazón está en reposo, lo que refleja la resistencia vascular y la elasticidad arterial. Esta relación fisiológica intrínseca explica la alta correlación observada entre ambas variables.[5]

Se exploraron tecnicas como la deteccion de datos atipicos con las tecnicas de IQR y Z-Score en la cual destaco el IQR en el context de detectar o acotar secciones de atipicos que se ajustan a las necesidades de este problema.

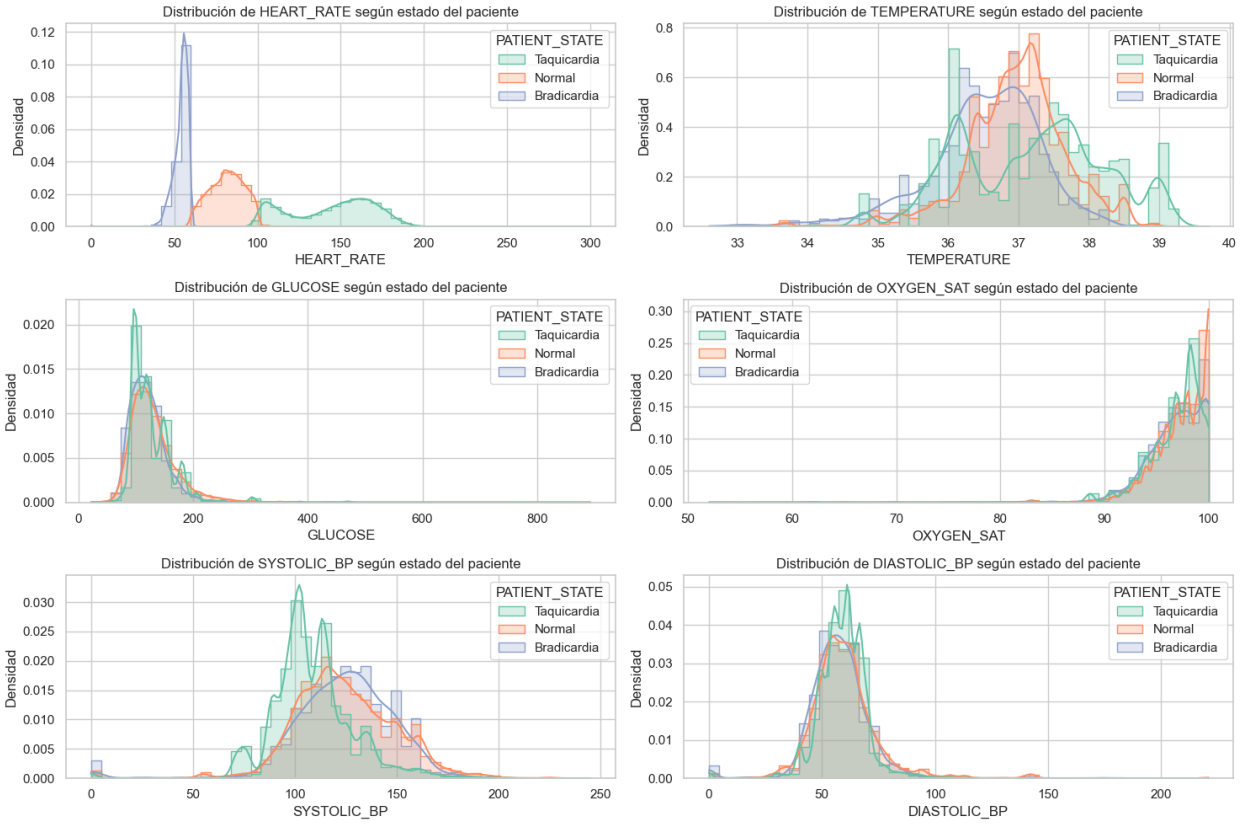


Figura 1. Distribucion de datos por categoria por signo vital

Pero el dataset presenta carencias como algunas secciones carentes de datos por lo cual para normalizer y estandarizar el dataset y dar tratamiento a los datos faltantes se decidio hacer un trabajo de limpieza, se empezo por estandarizar el formato de datos los cuales tuvieran registros tipo calendario para asegurar que la columna CHARTTIME esté en formato datetime “DD:MM:AA HH:MM:SS:MS”, en esta etapa se implementa metodos para agrupar y rellenar usando tecnicas de backward y forward evitar NaN dentro de misma hora.

Tambien se definen umbrales mínimos aceptables para algunas variables fisiológicas, con el objetivo de eliminar registros que, por su valor, representarían condiciones incompatibles con la vida o errores de medición evidentes. Para la temperatura corporal, de acuerdo con fuentes médicas, se considera que el cuerpo humano entra en un estado de hipotermia cuando la temperatura central desciende por debajo de los 35 °C. Temperaturas menores a este valor comprometen de manera crítica las funciones vitales, y valores inferiores a 20 °C serían prácticamente incompatibles con la vida [6]. Por este motivo, se eliminarán los registros con valores de TEMPERATURA menores a 20 °C, asumiendo que representan errores instrumentales o mediciones atípicas no fisiológicas. Para la saturación de oxígeno; según la literatura médica, una saturación de oxígeno por debajo de 90 % se considera un signo de hipoxemia, lo que indica

niveles insuficientes de oxígeno en la sangre y, en casos prolongados, puede ser potencialmente letal. [7] Por lo tanto, se establece un umbral mínimo de 50 % para esta variable, eliminando cualquier registro con valores inferiores a dicho límite.

En conjunto, estos criterios permiten mantener únicamente observaciones que representen condiciones fisiológicas realistas, evitando que datos erróneos que afecten las siguientes etapas del análisis.

Para la imputacion de datos faltantes y en vista de que el mecanismo de ausencia corresponde a un MAR (Missing At Random) como se vio en la literatura de tipos de problemas en datos faltantes [8], no se eliminan las variables con alta proporción de valores faltantes, como TEMPERATURE, ya que presentan relevancia fisiológica en la interpretación de los signos vitales (por ejemplo, el aumento de temperatura suele correlacionar con el incremento del ritmo cardíaco)[8]. En lugar de ello, se implementa una imputación mediante el algoritmo KNN (K-Nearest Neighbors Imputer), que estima los valores faltantes a partir de pacientes con características fisiológicas similares.

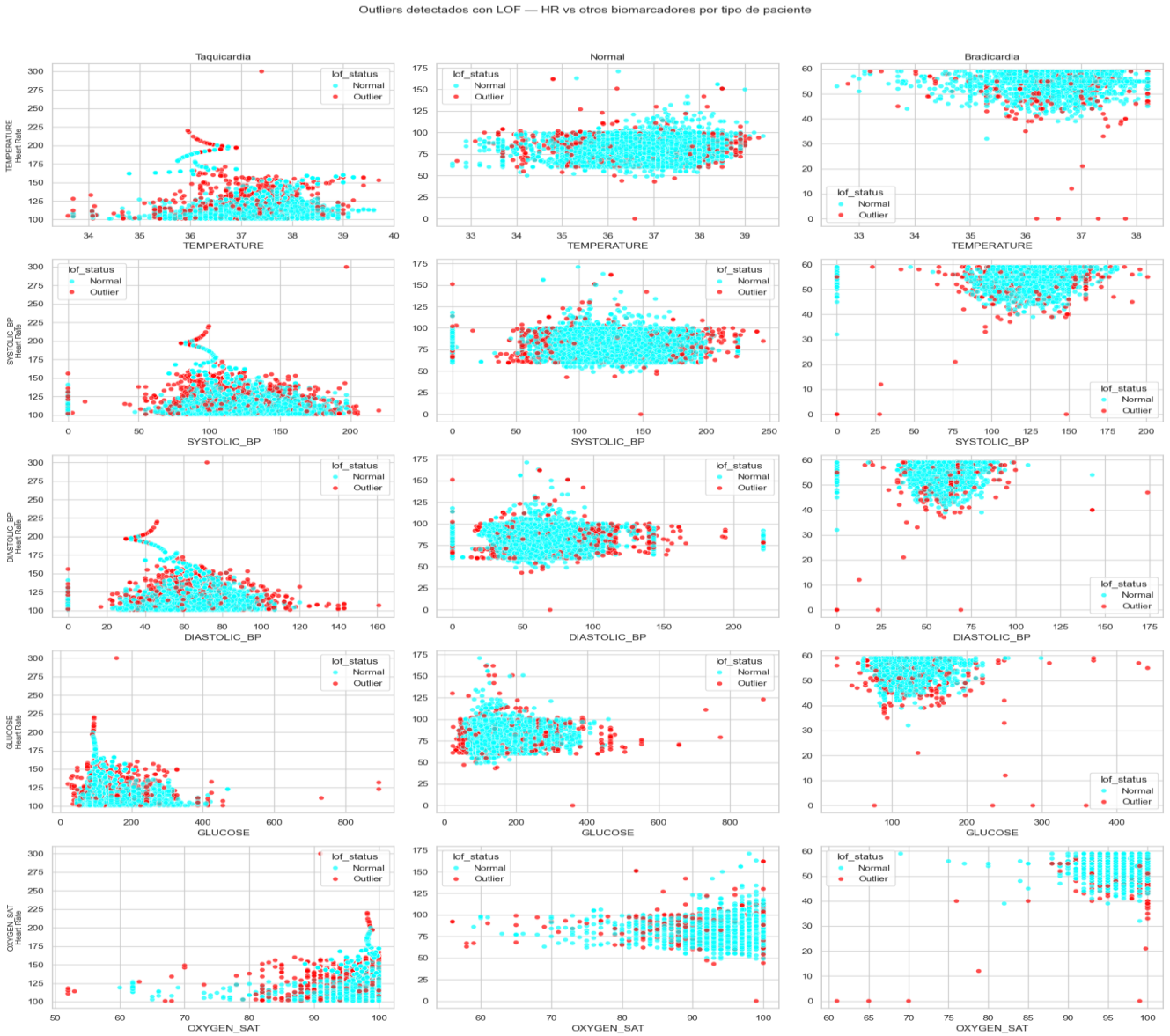
Este método tiene la ventaja de preservar las relaciones multivariadas entre los biomarcadores, reduciendo el sesgo introducido por imputaciones simples como la media o la mediana.

Con esto dimos finalizado la limpieza e imputacion de datos faltantes, quedando 72922 registros de los 100.000 acotados anteriormente; con esto se pudo dar rienda suelta al analisis de signos vitales por categoria y se encontro ciertas relaciones interesantes que notamos con el coeficiente de Spearman como unarelacion medianamente debil inversa entre la frecuencia cardiaca y la presion sistolica, lo cual puede explicarse por el reflejo barorreceptor, un mecanismo fisiológico mediante el cual el cuerpo regula la frecuencia cardíaca ante variaciones en la presión arterial, buscando mantener la estabilidad hemodinámica. [9] Se conserve tambien la correlación moderada entre la presión sistólica y la diastólica,

Por otro lado, la temperatura corporal presenta una correlación negativa débil con la saturación de oxígeno, lo que puede tener una explicación fisiológica relacionada con el efecto Bohr, en el que el aumento de la temperatura corporal disminuye la afinidad de la hemoglobina por el oxígeno, reduciendo ligeramente la saturación [10]

Finalmente se aplico una estrategia de deteccion de valores atipicos el cual era el objetivo principal de este Proyecto, para esto se uso Local Outlier Factor (LoF) y la aplicamos a cada una de las categorias con sus respectivas caracteristicas como se puede evidenciar en parte de la Figura 2

Figura 2. Deteccion de outliers con LoF



La tecnica LoF logra identificar varios valores que se desvían de los rangos fisiológicos esperados. En particular, en las variables de glucosa se destacan puntos con niveles significativamente altos, lo cual tiene sentido clínico, pues concentraciones elevadas de glucosa en sangre suelen considerarse anómalas en cualquier estado fisiológico. De igual forma, en la temperatura corporal, especialmente dentro de los grupos Normal y Bradicardia, se observan detecciones de outliers en valores alejados del rango corporal normal (aproximadamente entre 37,5 °C y 39.5 °C), lo que indica que es capaz de captar desviaciones térmicas extremas de manera coherente.

3. Results and Discussion

Al realizar un análisis exploratorio sobre la distribución de los signos vitales seleccionados, se enconto que las variables presentaban distribuciones similares entre las categorías “Normal”, “Taquicardia” y “Bradicardia”, con ligeras variaciones. En el análisis multivariado, solo se detectó una correlación fuerte entre la presión sistólica y la presión diastólica, coherente con la relación fisiológica entre ambas presiones como fases del ciclo cardiaco.

Para la detección de valores atípicos, al aplicar técnicas como el rango intercuartílico (IQR), Z-Score y el método Local Outlier Factor (LoF), entre estas LoF destacó por su capacidad para identificar anomalías consistentes aun que aun algo limitadas en la identificacion de varios grupos o subgrupos de outliers locales, el IQR demostro ser una Buena tecnica para detectar outliers aun que esta depende un poco mas del contexto y la calidad de los datos.

Se evidenciaron carencias en la calidad de los datos, con secciones incompletas o con valores faltantes que requirieron procesos de limpieza e imputación basados en técnicas de KNN para preservar relaciones fisiológicas entre variables.

Los resultados muestran que la metodología aplicada es una buena aproximación para la detección de outliers en signos vitales, permitiendo identificar desviaciones significativas que pueden alertar sobre condiciones clínicas críticas. No obstante, se reconoce la necesidad de mejorar las técnicas de limpieza y manipulación de datos para optimizar la precisión del análisis. Por otro lado, es critico contar con datos de mejor calidad y que incluyan todas las mediciones asociadas a los diferentes signos vitales para robustecer los modelos predictivos futuros.

Asimismo, se sugiere explorar nuevas técnicas y considerar correlaciones adicionales entre etiquetas específicas para grupos particulares de pacientes, de modo que el análisis sea más contextualizado y clínicamente relevante, al tener un analisis mas robusto.

4. Conclusiones

El monitoreo de signos vitales mediante modelos predictivos sera una herramienta muy útil para detectar a tiempo posibles problemas en pacientes, lo cual mejorara la atención hospitalaria y apoyara al personal de la salud reduciendo la carga y el estres en ambientes clinicos. Aunque las técnicas aplicadas funcionaron bien para identificar datos anómalos, todavía hay margen para mejorar la limpieza y el análisis de los datos para hacer modelos más precisos. Aun que se observaron patrones consistentes, como la fuerte correlación entre la presión sistólica y la diastólica,se debe mejorar y contar con un mejor dataset para que los resultados sean mas precisos y hechos a medida pues hubo limitaciones y se cree que se pudo llegar a mejores resultados de tener menos datos faltantes.

En resumen, este trabajo representa un paso importante hacia una monitorización más inteligente y automatizada, que puede ayudar a prevenir complicaciones y reducir costos hospitalarios. Además, se recomienda seguir perfeccionando las técnicas y buscar datos de mejor calidad para fortalecer futuros desarrollos

Bibliografia

[1] Cooper RJ, Schriger DL, Flaherty HL, Lin EJ, Hubbell KA. Efecto de los signos vitales en las decisiones de triaje. Ann Emerg Med. 2002 Mar; 39 (3):223-32. [PubMed]

[2] MÉNDEZ ORJUELA, Lida. ADVISE-BD: Modelo de detección de anomalías clínicas usando datos de los signos vitales de pacientes de UCIP de una cohorte del Hospital Militar Central de Colombia, mediante el uso de técnicas de aprendizaje automático. [en línea]. Universidad de los Andes, 2022153.

[3] Clinica Sagrado Corazon, *Protocolo de Monitoreo No Invasivo de Signos Vitales*, versión 002, abril 2013, código M-HO-G-010, 11 pp.

[4] Mayo Clinic, "Heart rate: What's normal?", *Mayo Clinic*, 2023. [Online]. Available: https://www.mayoclinic.org/healthy-lifestyle/fitness/expert-answers/heart-rate/faq-20057979

[5] J. A. Floras et al., "Variabilidad de la presión arterial y morbimortalidad: un análisis clínico," Rev. Esp. Cardiol., vol. 52, no. 12, pp. 1017-1031, 1999

[6] McIntyre L.A., Fergusson D.A., Hébert P.C., Moher D., Hutchinson J.S., "Prolonged therapeutic hypothermia after traumatic brain injury in adults. A systematic review," JAMA, vol. 289, no. 22, pp. 2992–2999, 2003.

[7] E. J. Topol, "Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again," in NCBI Bookshelf, National Center for Biotechnology Information (US), 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK482316. [Accessed: Nov. 25, 2025].

[8] Maria Bernarda Salazar, "Intro_data_2025", notebook de clase, Universidad de Antioquia, Medellin,2025.

[9]G. Mancia et al., “Baroreflex mechanisms in human cardiovascular regulation,” Circulation Research, vol. 116, no. 6, pp. 976–990, 2015. [Online]. Available: https://doi.org/10.1161/CIRCRESAHA.115.305374

[10] J. B. West, Respiratory Physiology: The Essentials, 10th ed. Philadelphia, PA: Wolters Kluwer, 2015.