Universidad Veracruzana

Facultad de Estadística e Informática

Región Xalapa

Especialización en Métodos Estadísticos

Modelos de clasificación para la detección de pacientes con sepsis

Reporte de aplicación
para obtener el diploma de Especialista en
Métodos Estadísticos

Presenta:
Karla Belen Ramírez Olivera

Director:
Dr. Saúl Domínguez Isidro

Codirectora:
Mtra. María Yesenia Zavaleta Sánchez

Asesor:
Mtro. Carlos Adrián Alarcón Rojas

Agosto de 2025

"Lis de Veracruz: Arte, Ciencia, Luz"

Universidad Veracruzana

Facultad de Estadística e Informática
Región Xalapa

Especialización en Métodos Estadísticos

Modelos de clasificación para la detección de pacientes con sepsis
Reporte de aplicación para obtener el diploma de Especialista en Métodos Estadísticos

Presenta:
LE. Karla Belen Ramírez Olivera

Director:
Dr. Saúl Domínguez Isidro

Codirectora:
Mtra. María Yesenia Zavaleta Sánchez

Asesor:
Mtro. Carlos Adrián Alarcón Rojas

UNIVERSIDAD VERACRUZANA

Facultad de Estadística e Informática

El Comité Académico de la Especialización en Métodos Estadísticos y el director de este trabajo recepcional intitulado: **Modelos de clasificación para la detección de pacientes con sepsis,** autorizan la impresión y la constitución del jurado para la defensa.

COMITÉ ACADÉMICO

Dra. Cecilia Cruz López
Coordinadora
Especialización en Métodos Estadísticos

Dr. Luis Gerardo Montané Jiménez
Director
Facultad de Estadística e Informática

MCIG. Emmanuel Morales García
Secretario

Dr. Ángel Juan Sánchez García
Vocal

Dra. Cecilia Cruz López
Presidente

GENERACIÓN: 2025 **SEDE: Xalapa**

TÍTULO: Modelos de clasificación para la detección de pacientes con sepsis

AUTOR: Karla Belen Ramírez Olivera

DIRECTOR: Dr. Saul Domínguez Isidro

CO-DIRECTOR: Mtra. María Yesenia Zavaleta Sánchez

TIPO DE TRABAJO:

Reporte de aplicación		Trabajo Practico-Educativo		Desarrollo estadístico		Monografía		Artículo o Capítulo de libro	
	X								

TIPO DE ESTUDIO:

Exploratorio	X	Retrospectivo	
Descriptivo	X	Prospectivo	
Expostfaco o cuasiexperimental		Transversal	
Experimental		Longitudinal	

METODOLOGÍA ESTADÍSTICA:

A) Diseño:

Muestreo	
Experimento	X
Estudio observacional	

B) Análisis

Exploratorio	X
Descriptivo básico	X
Inferencia básica	X
Métodos multivariados	
Regresión	X
ANOVA y ANCOVA	
Control de calidad	
Métodos no paramétricos	
Modelos especiales	
Técnicas avanzadas	X
Series de tiempo	

Dedicatoria

En memoria de mi papá, Ángel Ramírez Vidal y mi abuelito Ramón Ramírez Mora. Dos pilares fundamentales en mi vida, que, aunque ya no estén físicamente presentes, su recuerdo y enseñanzas siempre estarán conmigo.

A mi padre, por inculcarme los valores fundamentales, la perseverancia y la honestidad. Por ser mi guía y ejemplo por seguir en la vida. Te extraño cada día y sé que estas muy orgulloso de este logro.

A mi abuelito, por su sabiduría, sus consejos y por siempre cree en mi potencial. Gracias por ser mi segundo padre, por apoyar en mi educación. Tu presencia se hace sentir en cada paso que doy.

Y a mi madre, que sin sus grandes esfuerzos no hubiera hecho este sueño realidad, por acompañarme noches enteras, por buscar el sustento para podernos apoyar no solo a mi sino a mis hermanas también.

Con amor y eterna gratitud Karlita.

Tabla de contenido

Resumen

La sepsis representa una de las principales causas de mortalidad hospitalaria a nivel mundial, por lo que su detección temprana resulta esencial para mejorar el pronóstico de los pacientes. Este estudio tuvo como objetivo la aplicación de modelos de clasificación predictiva que permitan identificar factores clínicos relevantes asociados al desenlace de pacientes con sepsis. Para ello, se emplearon datos clínicos reales de 65 pacientes del Instituto Nacional de Cardiología Dr. Ignacio Chávez, y se implementaron dos enfoques analíticos: regresión logística y redes bayesianas. Ambos modelos fueron evaluados con y sin balanceo de clases mediante la técnica SMOTE, y validados a través de la validación cruzada. Los resultados indican que la regresión logística obtuvo un mejor desempeño en términos de precisión, mientras que la red bayesiana ofreció mayor interpretabilidad clínica. Se concluye que ambos modelos son útiles para apoyar la toma de decisiones médicas en contextos con recursos de datos limitados. Además, se proponen futuras líneas de investigación centradas en la reducción de dimensionalidad y el desarrollo de sistemas de alerta temprana.

Palabras clave: sepsis, modelos predictivos, regresión logística, red bayesiana, SMOTE, aprendizaje automático.

Capítulo 1. Introducción

La sepsis es una emergencia médica que describe la respuesta inmunológica sistémica del cuerpo a un proceso infeccioso, el cual puede conducir a la difusión orgánica y la muerte (Banchón Alvarado et al., 2020). A pesar de los avances en las investigaciones para la comprensión de la fisiopatología de este síndrome clínico, así como el surgimiento y mejoras en las herramientas de monitoreo hemodinámico y las medidas de reanimación, la sepsis sigue siendo una de las principales causas de morbilidad y mortalidad en pacientes críticos (Banchón Alvarado et al., 2020).

Las investigaciones en sepsis han sido abordadas desde los métodos estadísticos, centrándose en aspectos como la identificación de factores de riesgo, el análisis de patrones clínicos asociados y la evaluación de estrategias de tratamiento (Toro Beltrán 2022; Patel et al., 2023; Wang et al., 2023). En años recientes, el aprendizaje automático ha comenzado a integrarse en este campo, ofreciendo herramientas más sofisticadas para el análisis de grandes volúmenes de datos clínicos y permitiendo la creación de modelos predictivos más precisos (Kosyakovsky, 2022; Cai et al., 2024). No obstante, muchos de estos estudios enfrentan limitaciones, como la falta de interpretabilidad de los modelos, el sesgo en los conjuntos de datos utilizados o la falta de validación en contextos clínicos diversos, lo que deja abierta la necesidad de continuar explorando diferentes enfoques que permitan incrementar el conocimiento y el desarrollo tecnológico en esta área.

El desarrollo de esta investigación se basa en aplicar modelos de análisis predictivo para identificar factores críticos en la detección de sepsis temprana considerando un conjunto de datos con pocos registros, obtenidos de (Pale Carrion, 2006). A partir del conocimiento de los factores clínicos y la generación de dicho modelo, se busca contribuir a la mejora de la predicción y el manejo clínico de esta condición, impactando positivamente en la supervivencia de los pacientes.

1.1 Planteamiento del problema

Considerando el conjunto de datos bajo análisis procedentes del “Instituto Nacional de Cardiología Dr. Ignacio Chávez” de la Ciudad de México (Pale Carrion, 2006), se dispone de un conjunto limitado de registros clínicos, recolectados en un periodo de corta duración. Este escenario refleja una limitación común en muchas instituciones, donde los datos disponibles para investigaciones clínicas son escasos o no están completamente

estandarizados. Este escenario resalta la necesidad de desarrollar modelos predictivos capaces de trabajar con conjuntos de datos reducidos y específicos de contextos locales (Shaikhina **&** Khovanova, 2017).

Aunque las técnicas de aprendizaje automático han demostrado su potencial en la predicción de sepsis, muchos modelos existentes requieren grandes volúmenes de datos y carecen de aplicabilidad en entornos clínicos con recursos limitados. De acuerdo con lo presentado, ¿se puede desarrollar un modelo predictivo que no solo sea preciso, sino que también sea capaz de operar de manera efectiva con datos limitados y en un contexto clínico específico?

Este trabajo busca enfrentar el problema de generación de un modelo predictivo con alta precisión adaptado a un conjunto reducido, utilizando métodos estadísticos y aprendizaje automático. Se busca identificar factores críticos para la detección temprana de sepsis y establecer un precedente para integrar este modelo en sistemas de alerta temprana, mejorando así la capacidad de predicción y el manejo clínico en escenarios similares.

1.2 Preguntas de investigación

Las preguntas de investigación generadas para dar solución a esta investigación fueron las siguientes:

- ¿Cuáles son los factores clínicos más relevantes asociados con la detección temprana de sepsis en el contexto de los datos recolectados del "Instituto Nacional de Cardiología Dr. Ignacio Chávez"?
- ¿Un modelo de regresión logística generado en el contexto de los datos recolectados del "Instituto Nacional de Cardiología Dr. Ignacio Chávez", permite generar predicciones de sepsis con alta precisión, sensibilidad y especificidad?
- ¿Se puede generar una red bayesiana que muestre relaciones causales de factores clínicos asociados con la detección de sepsis en el contexto de los datos recolectados del "Instituto Nacional de Cardiología Dr. Ignacio Chávez"?

1.3 Objetivos

General

Analizar dos estrategias capaces de identificar factores asociados con la detección temprana de sepsis en pacientes hospitalizados, utilizando un conjunto limitado de datos clínicos provenientes del "Instituto Nacional de Cardiología Dr. Ignacio Chávez", asegurando precisión, sensibilidad y especificidad.

Particulares

- Caracterizar los factores clínicos más relevantes asociados con la detección temprana de sepsis en el contexto de un conjunto limitado de datos.
- Construir dos modelos predictivos utilizando técnicas de aprendizaje automático adaptadas para trabajar con conjuntos de datos reducidos.
- Validar los dos modelos predictivos en términos de precisión, sensibilidad, especificidad y área bajo la curva considerando las limitaciones inherentes del conjunto de datos.

1.4 Hipótesis

El uso de técnicas de aprendizaje automático logra identificar patrones clínicos relevantes en conjuntos de datos limitados, promoviendo la detección de sepsis con altos niveles de precisión, sensibilidad y especificidad, se valida su viabilidad para identificar factores críticos asociados con la detección temprana de esta condición.

Variables independientes (VI):

1. Técnicas de aprendizaje automático seleccionadas para el modelado.
2. Características del conjunto de datos como edad, tipo de infección, hemograma, perfil bioquímico y pruebas metabólicas básicas.

Variables dependientes (VD):

1. Desempeño del modelo predictivo, medido a través de precisión, sensibilidad, especificidad y AUC-ROC.
2. Capacidad del modelo para identificar factores clínicos relevantes asociados con sepsis temprana.

1.5 Justificación

Actualmente, muchos hospitales carecen de sistemas eficaces para el reconocimiento precoz de la sepsis, lo que resulta en diagnósticos tardíos y tratamientos inadecuados. Esto contribuye a altas tasas de mortalidad y costos elevados de atención médica. La utilización de datos clínicos para la predicción de sepsis podría ofrecer una solución prometedora a este problema. La identificación temprana y precisa de la sepsis es crucial para mejorar las tasas de supervivencia de los pacientes hospitalizados.

Ante este panorama, el desarrollo de un modelo de análisis predictivo fundamentado en técnicas de aprendizaje automático, utilizando únicamente un conjunto limitado de datos, tiene el potencial de transformar la práctica médica en dicho contexto. El beneficio directo de esta investigación es la posibilidad de detectar de forma anticipada los factores clínicos críticos asociados a la sepsis, con una base empírica sólida, facilitando la toma de decisiones medicas más rápidas y precisas.

Los principales beneficiarios directos de este modelo serán los profesionales médicos que contarán con una herramienta complementaria de apoyo clínico para identificar oportunamente pacientes en riesgo, lo que redundará en un mejor pronóstico y mayores tasas de supervivencia. A mediano plazo, los pacientes hospitalizados también se verán beneficiados al recibir intervenciones tempranas que eviten el desarrollo de complicaciones graves. De manera indirecta, el sistema hospitalario también se beneficiará mediante una optimización de recursos, al priorizar la atención a quienes realmente lo requieren, reduciendo costos derivados de cuidados intensivos prolongados y estancias hospitalarias innecesarias.

Finalmente, esta investigación también tiene un valor científico, al demostrar la viabilidad de aplicar técnicas de aprendizaje automático en escenarios con datos clínicos escasos o incompletos, abriendo la puerta a soluciones adaptadas a otros contextos hospitalarios similares, tanto en México como en otros países en vías de desarrollo.

Capítulo 2. Marco Teórico

Esta sección presenta el contexto del proyecto centrándose en establecer el entorno y las circunstancias específicas que rodean el problema planteado, las bases teóricas que sustentan el trabajo y sus antecedentes.

2.1 Marco Contextual

2.1.1 Definición y conceptos básicos de sepsis

Se le conocen como cuadros clínicos de sepsis, sepsis severa y shock séptico cuando se presentan un conjunto de manifestaciones clínicas, que van desde fiebre y taquicardia hasta disfunción multiorgánica y falla circulatoria. Estos cuadros se observan cada vez más en las unidades de cuidados intensivos y su alta mortalidad resalta en la necesidad de intervenciones tempranas y efectivas (Estupiñán et al., 2016).

Según la World Health Organization: WHO (2024), la sepsis es una afección médica crítica que ocurre cuando el sistema inmunitario del cuerpo responde de manera exagerada a una infección, resultando en disfunción orgánica. Este proceso puede dañar los tejidos y órganos, y si no se trata de inmediato, puede causar un choque séptico, insuficiencia multiorgánica y potencialmente la muerte. Personas de todas las edades pueden sufrir sepsis, pero los grupos más vulnerables incluyen a los ancianos, niños pequeños, embarazadas y aquellos con problemas de salud preexistentes.

Frausto (1999) describe la sepsis grave como una condición caracterizada por hipotensión o signos de hipoperfusión periférica, tales como oliguria, alteraciones agudas del estado mental o acidosis láctica. Además, el choque séptico se presenta en pacientes con sepsis grave cuya hipotensión no mejora con la administración de líquidos por vía parenteral, acompañado de indicios de hipoperfusión periférica.

Otra definición que nos da un punto de vista distinto a los conceptos presentados anteriormente es el de Estupiñán et al. (2016), que definen la sepsis como una respuesta inflamatoria sistémica frente a una infección documentada o sospechada, la cual se manifiesta con síntomas como fiebre, taquicardia y alteraciones en el conteo de leucocitos. Al progresar, la sepsis puede llevar a una disfunción orgánica, aumentando así la gravedad del cuadro clínico. Los mismos autores presentan una definición de sepsis severa, esta condición representa un estado avanzado de sepsis donde, además de la respuesta inflamatoria, se observan signos de hipoperfusión tisular y disfunción orgánica, tales como hipotensión

persistente, acidosis láctica, oliguria y alteraciones del estado mental, lo cual indica un alto riesgo de mortalidad. Estos signos reflejan el grado de compromiso de los órganos y la necesidad de atención intensiva.

El shock séptico se define como una etapa avanzada de sepsis severa caracterizada por una hipotensión persistente que no responde a la administración adecuada de líquidos, requiriendo el uso de medicamentos vasopresores para mantener la presión arterial. Cuando esta situación dura más de una hora sin responder a tratamientos con líquidos o fármacos se clasifica como shock séptico refractario (Briceño, 2005).

2.1.2 Importancia y contexto clínico

Como lo menciona el sitio web de Salud y Medicina (2019), la sepsis es un problema crítico en el ámbito de la salud pública debido a su elevada incidencia, alta mortalidad y significativo impacto económico. En México, su prevalencia alcanza el 12.9 %, una cifra superior a la registrada en países como Estados Unidos y Reino Unido, e incluso el doble en comparación con otras naciones como Turquía. Esta diferencia pone de manifiesto las desigualdades en el diagnóstico oportuno y en el acceso a tratamientos eficaces, especialmente en entornos hospitalarios con recursos limitados (Baxter,2025). Por otra parte, en España, se registran anualmente 175,000 casos de sepsis, de los cuales 50,000 son graves, causando aproximadamente 17,000 muertes. Estas cifras reflejan la magnitud del problema y su relevancia en las unidades de cuidados intensivos, donde se concentra la atención de los casos más severos. La mortalidad asociada a la sepsis se incrementa en un 8% por cada hora de retraso en la administración del tratamiento adecuado, lo que subraya la necesidad de un diagnóstico temprano y una intervención rápida.

Además de las implicaciones clínicas, la sepsis representa una carga económica considerable para los sistemas de salud. En España, cada episodio se estima en costos que oscilan entre 10,000 y 18,000 euros, lo que resalta la importancia de optimizar los recursos y los tiempos de atención. Estudios han mostrado que el retraso en el diagnóstico no solo aumenta los costos hospitalarios, sino también la probabilidad de complicaciones a largo plazo y mortalidad, aspectos que agravan el problema desde una perspectiva sanitaria y económica

A pesar de los avances en biomarcadores como la procalcitonina (PCT, por sus siglas en inglés) y la proteína C reactiva (CRP, por sus siglas en inglés), no existe aún un marcador único que permita un diagnóstico preciso por sí solo. Esto ha impulsado investigaciones en

nuevos indicadores como el MDW (ancho de distribución de monocitos, por sus siglas en inglés.), el cual podría mejorar la identificación temprana de pacientes con sepsis y permitir ajustes en el tratamiento basados en datos más específicos (Medicina, 2019).

2.1.3 Contexto de datos clínicos

Los datos se obtuvieron de los expedientes del Instituto Nacional de Cardiología Dr. Ignacio Chávez en la Ciudad de México, incluyendo notas médicas, historias clínicas y resultados de exámenes de laboratorio y gabinete, de pacientes adultos con enfermedades cardiacas, atendidos entre enero y agosto de 2006 (Pale Carrión, 2006).

Muestra: 65 pacientes hospitalizados que cumplieron con los criterios de inclusión:

- Criterios de inclusión utilizados:
  - Pacientes mayores de 18 años.
  - Diagnóstico confirmado de sepsis asociado a infecciones como neumonía, IVUS (infección de vías urinarias), celulitis, mediastinitis o bacteriemia.
- Criterios de exclusión utilizados:
  - Pacientes que no cumplan con las definiciones clínicas de las infecciones mencionadas.
  - Pacientes embarazadas.
  - Aquellos bajo tratamiento inmunosupresor, con neutropenia no asociada a sepsis o antecedentes de neoplasia.

Las características de los datos se presentan en la **Tabla 1,** las cuales se agrupan por 5 categorías, que incluye i) datos generales, ii) infecciones del paciente, iii) análisis de sangre completo (hemograma), iv) pruebas metabólicas básicas y v) perfil bioquímico. La mayoría de los datos son numéricos y solo el tipo de infección y el estado del paciente son categóricas.

Tabla 1. Datos clínicos recolectados de los pacientes			
Categoría	**Variable**	**Nombres clínicos**	**Tipo**
DATO GENERAL	EDAD	Edad del paciente en años	Numérica
INFECCIÓN	TIPO INF	Tipo de infecciones	Categórica
HEMOGRAMA	HTO	Hematocrito	Numérica
	LEUCOS	Leucocitos	Numérica
	PLAQ	Plaquetas	Numérica
PRUEBAS METABÓLICAS BÁSICAS	GLUC	Glucosa	Numérica
	BUN	Nitrógeno ureico en sangre (Blood Urea Nitrogen).	Numérica
	CREAT	Creatinina	Numérica
PERFIL BIOQUÍMICO	BILIS	Bilis	Numérica
	ALB	Albúmina	Numérica
	COL T	Colesterol total	Numérica
	TRIG	Triglicéridos	Numérica
	COL HDL	Colesterol bueno	Numérica
	COL LDL	Colesterol malo	Numérica
	PCR ING	Proteína C Reactiva al ingreso	Numérica
	PCR 24 HRS	Proteína C Reactiva a las 24 horas	Numérica
	PCR EGRE	Proteína C Reactiva al egreso	Numérica
TARGET	ESTADO_PAC	Sobrevivió o no sobrevivió	Categórica
Referencia: Elaboración propia			

2.2 Diagnóstico y detección de Sepsis

En este apartado se describe los principales enfoques clínicos y metodológicos utilizados para identificar la sepsis en pacientes. Son los criterios convencionales empleados por la práctica médica.

2.2.1 Detección tradicional de la Sepsis

Briceño (2005) menciona que para diagnosticar el Síndrome de Respuesta Inflamatoria Sistémica (SIRS o sepsis, deben estar presente dos o más de las siguientes condiciones o criterios:

1. Temperatura corporal superior a 38°C o inferior a 36°C.
2. Frecuencia cardiaca mayor de 90 latidos por minuto.
3. Frecuencia respiratoria mayor a 20 respiraciones por minuto o presión parcial de dióxido de carbono (PaCO2) inferior a 32 mmHg.
4. Y un recuento de leucocitos mayor a 12000 por mm3, o con más del 10 % de formas inmaduras.

Los grupos de riesgo pueden ser personas con una infección, una lesión o una enfermedad grave no transmitible puede desarrollarse en sepsis, el riesgo es más elevado en poblaciones vulnerables como las siguientes: personas mayores; mujeres embarazadas o que han dado a luz recientemente; recién nacidos; pacientes hospitalizados; pacientes en unidades de cuidados intensivos; personas con sistemas inmunitarios debilitados por ejemplo, debido al VIH (Virus de Inmunodeficiencia Humana) o al cáncer; personas con enfermedades crónicas (renal o cirrosis) (WHO, 2024).

Aunque las infecciones bacterianas son la causa más fuerte, la sepsis también puede originarse por infecciones virales, parasitarias o fúngicas.

El tratamiento de la sepsis requiere atención médica inmediata, incluyendo el uso de antimicrobianos y la admisión de líquidos por vía intravenosa. Los patógenos resistentes a los medicamentos representan un desafío significativo, ya que pueden llevar rápidamente a sepsis y choque séptico incrementando el riesgo de mortalidad. Para prevenir la sepsis, es esencial implementar medidas preventivas, como la higiene adecuada de las manos, acceso a programas de vacunación, y la mejora de sistemas de saneamiento y abastecimiento de agua. Un diagnóstico temprano y una atención médica adecuada y oportuna, incluyendo el uso óptimo de antimicrobianos y la rehidratación, son cruciales para mejorar las probabilidades de supervivencia. La sepsis puede tener un inicio repentino y ser mortal a corto plazo, pero también puede causar morbilidad significativa a largo plazo, requiriendo un enfoque multidisciplinario en su tratamiento (WHO, 2024).

2.3 Aprendizaje Máquina

El Aprendizaje Máquina es una rama de la Inteligencia Artificial que permite a las computadoras aprender de los datos sin necesidad de ser programadas explícitamente para cada caso. En lugar de que un desarrollador describa instrucciones para todas las situaciones posibles, el algoritmo aprende a partir de grandes volúmenes de información, identificando patrones y comportamientos. Este proceso le permite tomar decisiones o hacer predicciones con base en lo que ha aprendido previamente. Existen dos tipos principales de aprendizaje: el supervisado, donde se entrena al modelo con ejemplos que incluyen tanto las características como las respuestas esperadas, y el no supervisado, donde el modelo debe encontrar patrones sin conocer las respuestas de antemano. Dentro del aprendizaje supervisado se distingue principalmente dos tipos de algoritmos: los de clasificación y los de regresión (Sandoval Serrano, 2018).

2.3.1 Redes bayesianas

Las redes bayesianas son modelos gráficos probabilísticos que permiten representar de forma estructurada a las relaciones de dependencia entre un conjunto de variables aleatorias. Se definen como grafos dirigidos a cíclicos (DAG, por sus siglas en ingles), donde cada nodo representa una variable y los arcos dirigidos indican una relación de dependencia condicional. Esta estructura gráfica permite expresar la distribución de probabilidad conjunta de las variables como el producto de distribuciones condicionales más simples facilitando su manipulación computacional (López de Castilla-Vásquez, 2005). Esto se expresa formalmente como:

$$P(X_1, X_2, \dots, X_n) = \prod_{1=1}^{n} P(X_i | Pa_i)$$

Donde Pa_i representa los padres de la variable X_i en la estructura de la red (Friedman et al., 1997).

El clasificador bayesiano ingenuo es una forma simplificada de red bayesiana en la que se asume que todos los atributos son condicionalmente independientes entre sí, dado el valor de su clase. Su estructura gráfica consiste en un nodo raíz que representa la clase y nodos hijos que representan los atributos. Bajo esta configuración, la probabilidad de la clase C dada una instancia con atributos $A_1, A_2, \dots, A_n$ se calcula como:

$$P(C|A_1, \dots, A_n) \propto P(C) \prod_{1=1}^{n} P(A_i|C)$$

Aunque esta suposición es poco realista en muchos escenarios, este modelo ha demostrado ser sorprendentemente efectivo en tareas de clasificación (Friedman et al., 1997).

El entrenamiento de estas redes se basa en maximizar funciones de puntuación como la Longitud Mínima de Descripción (MDL), que equilibra la calidad del ajuste a los datos con la complejidad del modelo:

$$MDL(B|D) = \frac{log\, N}{2}|B| - LL(B|D)$$

donde $|B|$ es el número de parámetros del modelo, N es el tamaño del conjunto de datos, y $LL(B|D)$ es el algoritmo de la verosimilitud del modelo respecto a los datos. Este enfoque ayuda a evitar el sobreajuste penalizando modelos excesivamente complejos (Friedman et al., 1997). Además, para evitar estimaciones poco fiables debido a datos escasos, se emplea un suavizado bayesiano utilizando priors de Dirichlet:

$$P(X = x|D) = \frac{N}{N + N_0}\hat{P}_D(x) + \frac{N_0}{N + N_0}\theta_0(x)$$

donde N es el número de muestras observadas, N_0 es el peso del prior, y $\theta_0(x)$ es la probabilidad a priori del evento.

2.3.2 Regresión logística

La Regresión Logística es una técnica estadística multivariada tanto con fines explicativos como predictivos. Su principal característica es que está diseñada para aplicarse cuando la variable dependiente es dicotómica, es decir, presenta únicamente dos posibles estados (por ejemplo: si/no, pobre/no pobre, enfermo/no enfermo), los cuales suelen codificarse como 0 y 1. Las variables independientes, en cambio, pueden ser tanto cuantitativas como categóricas, aunque en este último caso debe transformarse en variables dicotómicas denominadas dummies (Chitarroni, 2002).

El propósito fundamental del modelo es estimar la probabilidad de que ocurra un evento específico en función de un conjunto de predictores. Por ejemplo, se puede calcular la probabilidad de que una persona sea pobre dependiendo de factores como el nivel educativo, la cantidad de personas ocupadas en el hogar o la edad promedio del grupo familiar.

La forma matemática de la expresión logística se basa en el modelo de los odds (razones de probabilidad), expresado así:

$$\frac{P}{1-p} = e^{z}$$

Donde P es la probabilidad de que ocurra el evento, e es la base del logaritmo natural, y Z es la combinación lineal de las variables independientes:

$$Z = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

Donde a es la contante del modelo (intercepto), b_1 son los coeficientes de regresión para cada variable X_i y por último, X_i son las variables independientes (predictoras).

También puede expresarse como:

$$P = \frac{e^{z}}{1+e^{z}} = \frac{1}{1+e^{-z}}$$

Es conocida como la función logística o sigmoide, y transforma cualquier valor de Z (que puede ir de $-\infty\ a\ +\infty$) a un rango de posibilidades entre 0 y 1.

Un aspecto importante del modelo es su capacidad de clasificación. A cada individuo se le asigna una probabilidad predicha de pertenecer a la categoría 1 (evento), y se clasifica en dicha categoría si esta probabilidad supera un cierto umbral (por lo general, 0.5). La calidad del ajuste se evalúa mediante estadísticas como el logaritmo de verosimilitud y tablas de clasificación, que muestran la proporción de casos correctamente predichos (Chitarroni, 2002).

A su vez, la regresión logística también permite transformar variables categóricas en indicadores binarios (dummies), lo que amplía su aplicabilidad a situaciones en que los predictores no son exactamente numéricos. Por ejemplo, una variable con cuatro categorías puede representarse mediante tres variables dicotómicas (dejando una categoría base), lo cual permite incluir factores cualitativos en el análisis (Chitarroni, 2002).

2.4 Antecedentes

El presente trabajo parte del estudio de Pale Carrion (2006), el cual se centró en determinar las características clínicas y hemodinámicas de la sepsis en pacientes cardiópatas. Además, buscaba establecer las diferencias entre los pacientes que sobreviven y aquellos que no, para mejorar diagnósticos en este grupo específico de pacientes. Emplearon técnicas estadísticas descriptivas, correlación de Pearson para identificar aquellas relaciones significativas, prueba T de Student para comparar grupos (sobreviviente y no sobrevivientes) y análisis multivariado para identificar variables significativas entre ambos grupos. Los resultados mostraron que la sepsis es más común en hombres (62.2 %) y la causa principal fue neumonía

(46.3 %). Se logró identificar que 35 de los pacientes fallecieron por factores como hipertensión arterial, cardiopatía isquémica y diabetes mellitus, estos asociados a la mortalidad (Pale Carrion, 2006).

2.4.1 Trabajos relacionados

En México Sierra Juárez et al. (2024), llevaron a cabo una validación de un modelo de inteligencia artificial para predecir el pronóstico de mortalidad en pacientes hospitalizados con sepsis. En dicho estudio, el modelo de redes neuronales fue el más efectivo, alcanzando un AUC (Área Bajo la Curva, por sus siglas en inglés) de 0.80 en la fase de adiestramiento y prueba. Sin embargo, los resultados finales mostraron un AUC promedio de 0.795 durante la validación cruzada. Los modelos de bosques aleatorios y máquinas de soporte de vectores obtuvieron AUC de 0.667 y 0.641 respectivamente, lo que sugiere que estos modelos no fueron tan efectivos como las redes neuronales en la predicción de la mortalidad. El modelo de redes neuronales tiene potencial, pero requiere de más datos para mejorar su rendimiento, ya que la muestra utilizada fue pequeña y no estaba equilibrada (más pacientes no fallecieron que los que sí lo hicieron).

Otro estudio realizado en Estados Unidos por Song et al. (2021), tenían como objetivo identificar los factores clínicos que se asocian con un tratamiento rápido de la sepsis en pacientes adultos, implementaron un modelo de Aprendizaje Máquina (Gradient Boosting Machine) para analizar datos de pacientes en urgencias, aplicando técnicas de selección de características para identificar los factores más importantes. Los factores como la presión arterial mínima, la frecuencia cardíaca inicial y los valores de la Escala de Coma de Glasgow (GCS, por sus siglas en inglés) fueron identificados como influyentes en el tratamiento rápido. Además, se observó que los modelos lograron altos niveles de precisión en la predicción del tratamiento rápido en dos subgrupos de pacientes.

Borges Sa (2024), llevo a cabo una investigación en todas las áreas de hospitalización, Urgencias y de UCI (Unidad de Cuidados Intensivos) del Hospital Universitario Son Llàtzer, en Palma de Mallorca, España con el propósito de desarrollar y validar modelos predictivos para la detección temprana de sepsis grave (SG) y shock séptico (SS) en pacientes mayores de 14 años utilizando técnicas de Big Data, Inteligencia Artificial y Aprendizaje Automático. Comparar su efectividad con métodos diagnósticos tradicionales. Los datos estructurados y no estructurados de la Historia Clínica Electrónica (HCE), como signos vitales, resultados de laboratorio, prescripciones y resúmenes de alta. Los datos no estructurados fueron

procesados mediante técnicas de Procesamiento de Lenguaje Natural (PLN). Dicho análisis incluyó 815,170 registros de HCE, representando 461,392 episodios de pacientes. Los modelos predictivos se compararon con los métodos tradicionales, evaluando métricas como el AUC-ROC (curva de características operativas del receptor), sensibilidad, y especificidad. Finalmente, el mejor modelo predictivo fue un "ensemble" que combinaba Aprendizaje Automático con el criterio de Sepsis 2, obteniendo un AUC-ROC de 0.95, una sensibilidad del 93 %, y una especificidad del 84 % (Borges Sa, 2024).

Toro Beltrán (2022), presentó su trabajo de grado sobre información clínica de pacientes con sepsis, estructurada bajo el estándar HL7 (Health Level Seven, por sus siglas en inglés), FHIR (Fast Healthcare Interoperability Resources, por sus siglas en inglés) y CDA (Clinical Document Architecture, por sus siglas en inglés), para facilitar la visualización en un tablero de control (dashboard) que permita un diagnóstico oportuno. Su recolección de datos fue de pacientes hospitalizados de diversas instituciones de salud. Los datos incluyen variables como signos vitales, resultados de laboratorio y datos demográficos. Dichos datos clínicos fueron estandarizados utilizando HL7 FHIR, que permite integrar información desde múltiples fuentes de salud en un formato estructurado y compatible. Los datos fueron analizados mediante herramientas como MongoDB y Python. Se añadieron variables de ingeniería para enriquecer el análisis. Desarrollo modelos de Aprendizaje Automático que evaluaron estas variables para predecir la aparición de sepsis. Los modelos predictivos aplicados en el proyecto ofrecieron un diagnóstico oportuno con alta precisión, proporcionando al personal médico una herramienta efectiva para la toma de decisiones.

El estudio de Sierra Juárez et al. (2024), se basó en el uso de Inteligencia Artificial para mejorar la predicción de la mortalidad en pacientes con sepsis, en comparación con las escalas tradicionales como SOFA (Sequential Organ Failure Assessment), qSOFA (Quick Sequential Organ Failure Assessment), APACHE II (Acute Physiology and Chronic Health Evaluation II), entre otras. El tipo de investigación fue observacional que incluyó 218 expedientes electrónicos de pacientes adultos en el Hospital Central del Estado de Chihuahua, México, de julio de 2018 a enero de 2022. Aplicaron tres algoritmos: redes neuronales, máquinas de soporte vectorial y bosques aleatorios. El modelo de redes neuronales obtuvo el mejor desempeño, con un área bajo la curva promedio de 0.795 tras validación cruzada, aunque ligeramente inferior al valor deseado (>0.80), fue superior a los otros modelos. Los autores concluyeron que este tipo de modelos tienen potencial para ser implementado en

entornos hospitalarios, siempre que se amplíe la muestra y se mejore el balance de clases, permitiendo así una predicción más precisa del riesgo de mortalidad durante las primeras 24 horas de estancia intrahospitalaria.

Ríos Bolaños (2022), desarrolló un algoritmo que facilita la detección temprana de la sepsis a partir de un conjunto de datos clínicos, con el fin de predecir su aparición. Implementó la aplicación de técnicas de Aprendizaje Automático, destacando el uso de clasificadores como el Bayesiano, Máquinas de Soporte Vectorial, y la Regresión Logística y lineal. De los clasificadores aplicados, el Bayesiano resultó ser el más efectivo. Las estrategias de combinación de tipo producto y votación por mayoría también se destacaron. Los mejores resultados se obtuvieron utilizando una ventana de tiempo deslizante de 8 horas.

El estudio desarrollado por Chicco y Jurman, (2020), tuvo como objetivo predecir la supervivencia de los pacientes con sepsis utilizando solo tres variables fácilmente disponibles: edad, sexo y número de episodios sépticos. Para ello, crearon modelos predictivos rápidos y efectivos utilizando algoritmos de aprendizaje automático para mejorar la toma de decisiones clínicas. El estudio fue de tipo cuantitativo basado en análisis de cohortes. Se emplearon datos de 110,204 hospitalizaciones en Noruega y una cohorte de validación externa de 137 pacientes en Corea del Sur. Se implementó el uso Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés), Regresión Lineal, Potenciación del Gradiente y Ingenuo Bayes, aplicados a datos clínicos mínimos (edad, sexo y número de episodios sépticos) para predecir la supervivencia de los pacientes. Se concluyó que los mejores modelos lograron áreas bajo la curva de precisión-recuperación de hasta 0.966 en la cohorte noruega y 0.863 en la cohorte de validación surcoreana. Estos resultados sugieren que es posible predecir la supervivencia en sepsis con un conjunto mínimo de datos clínicos.

Capítulo 3. Metodología

En este capítulo se detalla el tipo de estudio, los criterios de inclusión y exclusión, las etapas de tratamiento de datos, así como los procedimientos utilizados para construir y evaluar los modelos predictivos.

3.1 Tipo de estudio

Este estudio descriptivo y exploratorio con un enfoque cuantitativo, dado que se utilizarán técnicas estadísticas y de aprendizaje automático para analizar datos clínicos de pacientes con sepsis. En particular, se empleará un modelo de red bayesiana con fines predictivos y explicativos.

- Descriptivo: Se analizarán los factores clínicos que influyen en la detección de la sepsis.
- Exploratorio: Debido al uso de datos clínicos limitados y la necesidad de explorar patrones mediante técnicas de aprendizaje automático y métodos estadísticos avanzados, como análisis multivariado o modelado estadístico.
- Inferencial: debido a que se generan predicciones para validar los modelos.

3.2 Justificación del modelo bayesiano

Se optó por el uso de redes bayesianas debido a su capacidad para modelar relaciones probabilísticas y dependencias condicionales entre múltiples variables clínicas, incluso cuando se dispone de conjuntos de datos pequeños o incompletos. Además, permiten integrar conocimiento previo y ofrecen una visualización comprensible del comportamiento de las variables frente a un desenlace binario, como la supervivencia o el fallecimiento del paciente. Esto resulta especialmente útil en contextos clínicos donde la incertidumbre es alta. Y las decisiones deben ser rápidas y fundamentadas.

3.3 Población objetivo

La población objetivo incluye pacientes hospitalizados con sospecha o diagnóstico confirmado de sepsis en el "Instituto Nacional de Cardiología Dr. Ignacio Chávez", Ciudad de México.

- Criterios de inclusión:

- Pacientes con datos clínicos completos registrados durante el período de estudio.
- Pacientes mayores de 18 años.

- Criterios de exclusión:
  - Registros clínicos incompletos.
  - Pacientes con diagnósticos distintos a sepsis o sin sospecha documentada.

Dado que el estudio trabajará con un conjunto reducido de datos, se considerarán todas las muestras disponibles en el período de recolección de datos.

3.4 Diseño de análisis de datos

Análisis de datos: Análisis descriptivo de las variables.

Tratamiento de datos: Limpieza, estandarización y preprocesamiento de los datos para su uso en el modelo.

3.4.1 Desarrollo del modelo predictivo

Preprocesamiento y balanceo de clase: inicialmente se construyó un modelo bayesiano con los datos in balancear, pero los resultados no fueron tan relevantes, posiblemente al desequilibrio de las clases. Para corregir esta asimetría se aplicó la técnica SMOTE (Técnica de Sobremuestreo de Minorías Sintéticas) la cual genera nuevas instancias sintéticas basadas en los vecinos más cercanos (Moreno et al., 2009).

Modelado de la red bayesiana: Para ellos se construyó una red bayesiana empleando el algoritmo de búsqueda Asenso de Colina (Hill Climber). Este algoritmo permite aprender la estructura óptima del modelo directamente desde los datos, sin imponer restricciones innecesarias. Se construyó la red mediante el software Weka (Hall et al., 2009).

3.4.2 Modelado complementario: Regresión logística

Para complementar el análisis y comparar enfoques predictivos, se aplicó un modelo de regresión logística utilizando R (RCore Team, 2023). Esta técnica estima la probabilidad de un desenlace (fallecimiento o supervivencia) en función de múltiples predictores clínicos. Se utilizó la función glm() para el ajuste del modelo, y se analizaron los coeficientes, su significancia estadística (prueba de Wald) y los odds ratios asociados a cada predictor.

3.5 Evaluación del modelo

Para devolver el desempeño de los modelos predictivos, tanto la Red Bayesiana como la Regresión Logística se implementaron dos enfoques: con datos originales (sin balancear) y con datos balanceados utilizando la técnica de "Remuestreo un conjunto de datos aplicando

la Técnica de Sobremuestreo de Minorías Sintéticas (SMOTE, por sus siglas en inglés), el cual es un algoritmo de *oversampling* que genera instancias "sintéticas" o artificiales para equilibrar la muestra de datos basado en la regla del vecino más cercano (Moreno, et al., 2009).

En el caso de la Red Bayesiana se empleó en el software WEKA y se utilizó el método de validación cruzada estratificada de 10 iteraciones (10-fold cross-validation) para estimar la capacidad predictiva del modelo y reducir el riesgo de sobreajuste, dada la naturaleza limitada del conjunto de datos. Se evaluó el rendimiento tanto del modelo construido con los datos originales como del modelo construido tras aplicar SMOTE.

Por su parte, la Regresión Logística fue implementada en software R Studio. También se compararon los resultados obtenidos con los datos sin balancear y con los datos ajustado mediante SMOTE. Para ambos escenarios se utilizó la estrategia de validación cruzada de 10 pliegues, generando la representatividad de ambas clases (sobrevivientes y no sobrevivientes) en cada iteración.

Los modelos fueron evaluados mediante las siguientes métricas de desempeño:

- Precisión (accuracy): proporción total de clasificaciones correctas.
- Sensibilidad (recall para no sobrevivientes): capacidad del modelo para identificar correctamente a los pacientes que no sobrevivieron.
- Especificidad: proporción de pacientes sobrevivientes correctamente identificados.
- F1-puntaje: es una medida que se combina la precisión y la sensibilidad en un solo valor armónico, lo que permite evaluar el equilibrio entre los errores de tipo I (falsos positivos) y tipo 2 (falsos negativos).
- Área bajo la curva ROC (AUC, por sus siglas en inglés): medida global del poder discriminante del modelo.

Estas medidas fueron seleccionadas por su capacidad para reflejar distintos aspectos del rendimiento en problemas de clasificación binaria, especialmente en contextos clínicos donde la detección temprana de estos casos graves es prioritaria.

Capítulo 4. Resultados

En este capítulo se presentan los principales hallazgos derivados del análisis estadístico de los modelos predictivos. En primer lugar, se muestran los resultados descriptivos de las variables clínicas consideradas, lo que permite caracterizar la población de estudio. Posteriormente, se detallan los resultados obtenidos a partir de los modelos de clasificación, Regresión Logística y Red Bayesiana, tanto con los datos sin balancear como con los datos balanceados mediante la técnica SMOTE.

4.1 Análisis descriptivo

Se analizaron 65 pacientes con el diagnóstico de sepsis, de los cuales el 62.2 % (45) eran hombres y el restante eran mujeres (38.8 %) (Pale Carrión, 2006). El promedio de edad de los pacientes hospitalizados es de 55 años.

En la **Tabla 2** se observan los valores obtenidos para la media, desviación estándar, mediana, mínimo y máximo para cada variable por clase. Se muestra que la clase de pacientes que no sobreviven presentan medias y medianas mayores a los sobrevivientes en las variables EDAD, GLUC, BUN, CREAT, BILIS, TRIG, PCR ING, PCR 24 HRS y PCR EGRE.

Se observa que los pacientes que no sobrevivieron presentan en general valores mayores en variables como EDAD, GLUC, BUN, CREAT, BILIS, TRIGLICÉRIDOS y PCR (en ingreso, a las 24 horas y al egreso). Lo que sugiere que estos pacientes tenían una edad más avanzada y mayores alteraciones metabólicas, inflamatorias y renales al momento de su ingreso y durante su estancia hospitalaria. La glucosa es más alta en pacientes que no sobreviven. Los niveles de proteína C reactiva, son más consistentes en pacientes que no sobreviven en todas las mediciones. Y, por último, el nitrógeno ureico también es mayor, lo que puede indicar deterioro de la función renal.

Tabla 2. Estadísticas descriptivas por variable y por clase.

Variable	No sobrevive (N=24)					Sobrevive (N=41)				
	Media	Desv. estándar	Mediana	Min.	Max	Media	Desv. estándar	Mediana	Min.	Max
EDAD	59	10	60	37	78	53	14	53	21	85
HTO	34.75	8.35	31	21	49	34.92	6.03	35	26	52
LEUCOS	13.59	5.07	14.6	1.8	26	15.75	5.09	15.1	5.0	29
PLAQ	184.58	161.52	157	49	825	219.34	113.15	187	38	529
GLUC	165.08	83.11	137.5	64	380	149.34	57.39	141.0	71	293
BUN	46.33	29.39	39	12	133	30.73	22.70	25	10	125
CREAT	2.08	1.32	1.6	0.6	5.8	1.48	0.88	1.2	0.4	5.8
BILIS	1.66	1.80	1.04	0.4	8.0	1.28	0.85	1.00	0.3	3.2
ALB	2.85	0.65	2.8	1.9	4.50	3.19	0.56	3.1	2.2	4.39
COL T	111.95	54.76	104	51	280	130.92	35.84	141	56	210
TRIG	135.66	78.06	104	48	330	128.39	53.71	124	50	312
COL HDL	25.29	11.81	25	6	49	33.97	12.80	34	8	70
COL LDL	66.83	40.76	58.5	14	187	75.12	25.20	73.0	15	123
PCR ING	143.79	83.15	123	19	335	106.92	76.05	88	6	336
PCR 24 HRS	146.66	77.36	123.5	29	324	109.68	85.63	89.0	6	422
PCR EGRE	122.45	74.43	116	16	284	43.53	66.26	30	7	422

Referencia: Elaboración propia

En la **Figura 1** se muestran los diagramas de cajas para cada una de las variables, separadas por clase (fallece y sobrevive). Se observa que las variables presentan escalas de medición heterogéneas. Se evidencian datos atípicos en varias de las variables, principalmente dentro del grupo de pacientes que sobreviven, como en las variables GLUC, PCR 24 HRS y PLAQ. La variable GLUC se destaca por su variabilidad entre las clases, con una mayor dispersión y presencia de valores atípicos en el grupo de fallecidos. Además, se observan diferencias notables en la distribución de los valores entre las clases en las variables de BUN, CREAT, PCR 24 HRS y PLAQ.

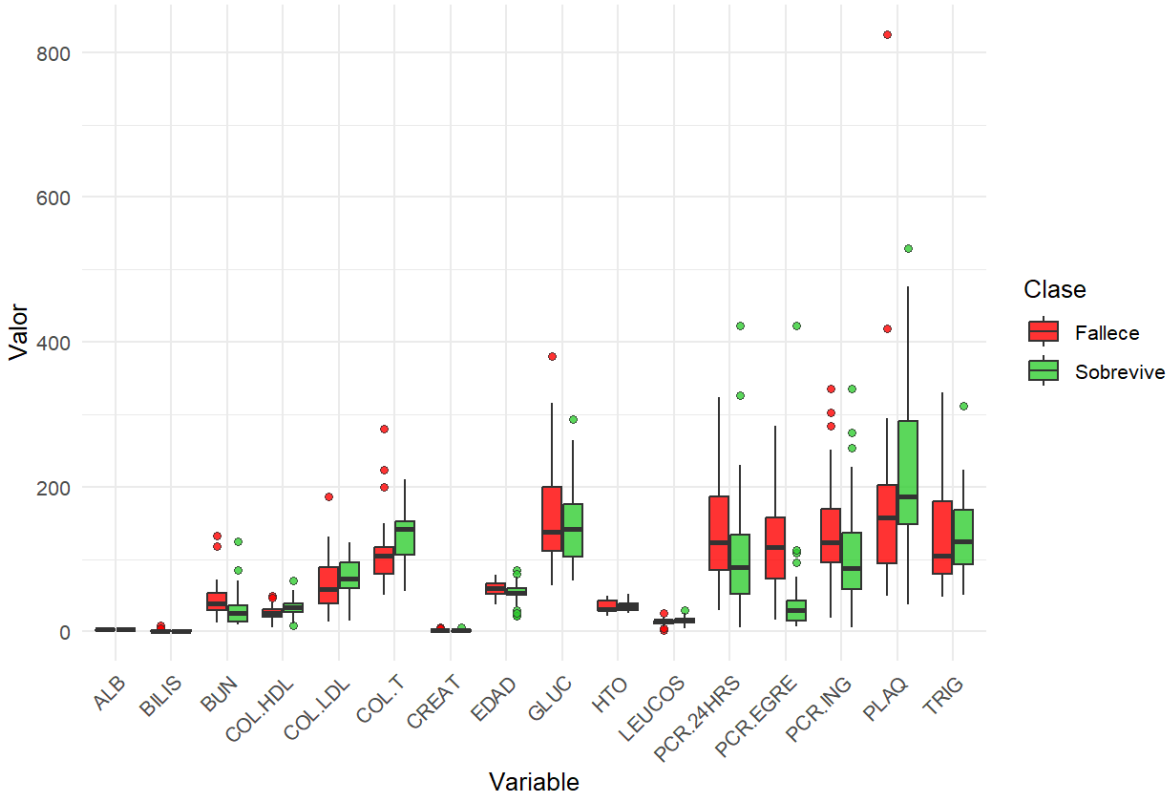

Figura 1. Gráfico de cajas para las variables por clase.

Referencia: Elaboración propia.

4.2 Análisis definitivo

En esta sección se presentan los resultados del análisis definitivo aplicado a los datos clínicos recopilados. Para ello se utilizaron dos enfoques estadísticos complementarios: la regresión logística y las redes de bayesianas. Ambos métodos permiten modelar la probabilidad de que un paciente con sepsis sobreviva o no, en función en diversas variables clínicas. La elección de estas técnicas responde tanto a su capacidad para manejar variables categóricas y continuas, como a su utilidad en contextos clínicos para la toma de decisiones médicas basadas en evidencia. A continuación, se detallan los procedimientos, resultados y consideraciones obtenidas para cada uno de los modelos implementados.

Con el fin de evaluar el impacto del desbalance de clases presentes en los datos, ambos modelos fueron ajustados en dos escenarios: utilizando el conjunto original de datos y posteriormente aplicando la técnica de *oversampling* para balancear la clase minoritaria. A

continuación, se presentan los procedimientos, resultados y comparaciones obtenidas en cada uno de los casos.

4.2.1 Regresión logística

La regresión logística permite estimar la probabilidad de ocurrencia de un evento binario, en este caso el estado del paciente si sobrevive o no en función de una o más variables independientes. Esta sección incluye el ajuste del modelo sin aplicar balanceo de clases y posteriormente los resultados obtenidos tras el uso de la técnica *oversampling*.

4.2.1.1 Regresión logística sin balanceo de clases

En esta primera etapa se ajustó un modelo de regresión logística utilizando el conjunto de datos original, es decir, sin aplicar ninguna técnica de balanceo. Esto permitió evaluar el comportamiento del modelo ante un desbalance natural en las clases donde la mayoría de los pacientes sobrevivieron. A continuación, se presentan los coeficientes estimados, métricas de desempeño y la capacidad predictiva del modelo bajo esta configuración.

El modelo de regresión logística se expresa en la siguiente ecuación.

$$P(Supervivencia) = \frac{e^{logit}}{1 + e^{logit}}$$

Donde:

$$\begin{aligned} logit = {} & 1.315 - 0.054 \cdot EDAD - 0.015 \cdot HTO + 0.199 \cdot LEUCOS - 0.0025 \cdot PLAQ \\ & - 0.00005 \cdot GLUC - 0.0367 \cdot BUN + 0.368 \cdot CREAT + 0.016 \cdot BILIS \\ & + 1.056 \cdot ALB - 0.0185 \cdot COL\_T + 0.0059 \cdot TRIG + 0.068 \cdot COL\_HDL \\ & - 0.0033 \cdot COL\_LDL - 0.0176 \cdot PCR\_ING + 0.0172 \cdot PCR\_24HRS \\ & - 0.0299 \cdot PCR\_EGRE \end{aligned}$$

En la **Tabla 3** se destaca que la cantidad de leucocitos (LEUCOS) y los niveles de proteína C reactiva al egreso (PCR EGRE) son predictores significativos para determinar el estado del paciente, específicamente su probabilidad de supervivencia. En particular, un número mayor de leucocitos se asocia con una mayor probabilidad de supervivencia, mientras que niveles elevados de PCR al momento del egreso disminuyen dicha probabilidad. Por ejemplo, por cada unidad adicional de leucocitos, el riesgo de supervivencia se

incrementa significativamente ($OR > 1, p = 0.041$), mientras que por cada unidad adicional de PCR EGRE, la probabilidad de sobrevivir disminuye ($OR < 1, p = 0.003$).

De igual forma los valores intermedios de proteína C reactiva durante el ingreso (PCR ING) y a las 24 horas (PCR 24 HRS) también muestran una asociación marginal con el desenlace, lo que sugiere que la evolución de la respuesta inflamatoria del paciente puede ser un factor determinante para su recuperación o fallecimiento.

Las demás variables (EDAD, GLUC, CREAT, etc.) no resultaron estadísticamente significativas ($p > 0.05$), por lo que su efecto sobre la probabilidad de supervivencia no puede confirmarse con los datos disponibles.

Tabla 3. Regresión logística

Características	β	Error estándar	Z	P	Exp (β)	OR
Intercepto	1.315e+00	4.901e+00	0.268	0.78847	3.7239126	(0.000,75622.471)
EDAD	-5.418e-02	5.545e-02	-0.977	0.32851	0.9472632	(0.832, 1.041)
HTO	-1.530e-02	6.214e-02	-0.246	0.8055	0.9848160	(0.868, 1.114)
LEUCOS	1.991e-01	9.781e-02	2.036	0.04178	1.2203295	(1.020, 1.516)
PLAQ	-2.547e-03	3.227e-03	-0.789	0.42999	0.9974561	(0.991,1.004)
GLUC	-4.975e-05	6.171e-03	-0.008	0.99357	0.9999502	(0.988, 1.012)
BUN	-3.670e-02	2.925e-02	-1.255	0.20960	0.9639685	(0.904, 1.019)
CREAT	3.678e-01	6.181e-01	0.595	0.55177	1.4445688	(0.408, 5.139)
BILIS	1.641e-02	3.750e-01	0.044	0.96509	1.0165486	(0.477, 2.201)
ALB	1.056e+00	1.011e+00	1.045	0.29623	2.8752835	(0.410, 25.257)
COLT	-1.846e-02	2.670e-02	-0.692	0.48922	0.9817047	(0.924, 1.027)
TRIG	5.915e-03	1.032e-02	0.573	0.56639	1.0059321	(0.987, 1.029)
COL HDL	6.832e-02	6.237e-02	1.095	0.27333	1.0707068	(0.962, 1.238)
COL LDL	-3.306e-03	2.507e-02	-0.132	- 0.89508	0.9966995	(0.948, 1.049)
PCR ING	-1.764e-02	1.031e-02	-1.711	0.08703	0.9825116	(0.960, 1.001)
PCR 24 HRS	1.719e-02	9.941e-03	1.729	0.08383	1.0173348	(0.999, 1.040)
PCR EGRE	-2.994e-02	1.030e-02	-2.907	0.00365	0.9705005	(0.947, 0.987)

Nota: OR = Odds ratio. **Técnica**

Los valores de p en rojo indican significancia estadística (p < 0.01); los valores en verde corresponden a asociaciones marginalmente significativas (p < 0.10).

Referencia: Elaboración propia

El valor del Criterio de Información Akaike (AIC por sus siglas en inglés) fue de 84.75 lo que indica que el modelo logra un equilibrio adecuado entre precisión y complejidad.

El área bajo la curva mide la calidad general del modelo en términos de su capacidad para distinguir entre dos clases.

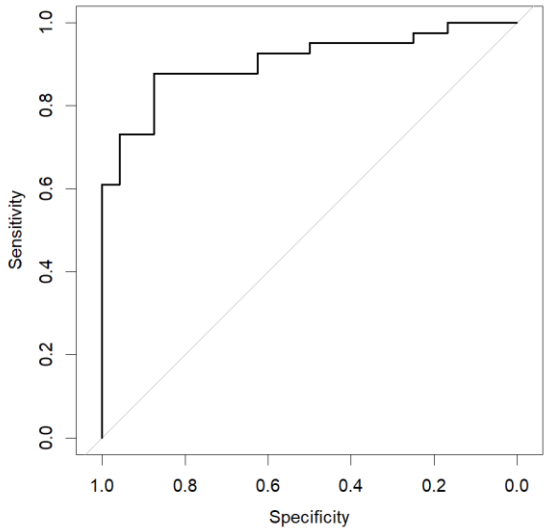

Figura 2. Área bajo la curva (sin balanceo)

Referencia: Elaboración propia.

Con el modelo de regresión logística se obtiene una capacidad global de predicción del estado del paciente del 81.5 %, con una predicción correcta de supervivencia del 73.2 % (30 verdaderos positivos de 41 casos reales de supervivencia), y una predicción correcta de fallecimiento del 95.8 % (23 verdaderos negativos de 24 casos reales de no supervivencia) (**Tabla 4)**. Esto indica que el modelo tiene un excelente desempeño para clasificar, lo cual se destaca por un área bajo la curva (AUC) de 0.9075, evidenciando una alta capacidad discriminativa del modelo (**Figura 2).**

Tabla 4. Matriz de confusión

Clase real	No sobrevivientes	Sobrevivientes
No sobrevivientes	23 (CC)	1 (FP)
Sobrevivientes	11 (FN)	30 (CC)

CC= Correctamente clasificados
FN= Falsos negativos

FP= Falsos positivos
Referencia: Elaboración propia

En **Tabla 5** se observa que el modelo reconoce bien a los pacientes que no sobreviven, con una sensibilidad del 95.8 %, lo que significa que identifica casi todos los casos reales de fallecimiento. Sin embargo, su precisión en esta clase es menor (67.6 %), lo que indica que también comete errores al predecir esta condición.

En cambio, en el caso de los pacientes sobrevivientes, el modelo es muy preciso (96.8 %), es decir, cuando predice que alguien sobrevivirá, lo hace casi siempre correctamente. No obstante, tiene una menor capacidad para detectar a todos los sobrevivientes (sensibilidad del 73.2 %), lo que significa que algunos son clasificados erróneamente como no sobrevivientes.

Tabla 5. Métricas por clase (sin balanceo)

Métricas	Sobrevivientes	No sobrevivientes
Presión	96.8 %	67.6 %
Sensibilidad	73.2 %	95.8 %
F1-Puntaje	83.3 %	79.0 %
Área bajo de la curva ROC	90.8 %	9.2 %

Referencia: Elaboración propia

4.2.1.2 Regresión logística con balanceo de clase

Posteriormente se aplicó la técnica SMOTE, un método de *oversampling* que genera instancias sintéticas de la clase minoritaria, con el propósito de balancear las clases y mitigar el sesgo provocado por la desproporción entre pacientes sobrevivientes y no sobrevivientes. Una vez equilibrado el conjunto de datos se procedió a reajustar el modelo de regresión logística y analizar sus resultados. Esta sección compara su desempeño con respecto al modelo previo sin balanceo, evaluando posibles mejoras en sensibilidad, especificidad y capacidad de clasificación global.

El modelo de regresión logística con e balaceo se expresa en la siguiente ecuación.

$$P(Supervivencia) = \frac{e^{logit}}{1 + e^{logit}}$$

Donde:

$$logit = -1.122 - 0.042 \cdot EDAD - 0.049 \cdot HTO + 0.240 \cdot LEUCOS - 0.001 \cdot PLAQ + 0.006 \cdot GLUC - 0.053 \cdot BUN + 0.627 \cdot CREAT + 0.363 \cdot BILIS + 1.069 \cdot ALB - 0.005 \cdot COL\,T + 0.006 \cdot TRIG + 0.082 \cdot COL\,HDL - 0.016 \cdot COL\,LDL - 0.021 \cdot PCR\,ING + 0.019 \cdot PCR\,24\,HRS - 0.030 \cdot PCR\,EGRE$$

En la **Tabla 6**, se observa que los leucocitos (LEUCOS), los niveles de nitrógeno ureico en sangre (BUN), la proteína C reactiva a las 24 horas (PCR 24 HRS) y al egreso (PCR EGRE) resultaron ser predictores estadísticamente significativos para determinar la probabilidad de supervivencia en pacientes con sepsis. En particular un mayor conteo de leucocitos se asocia con un aumento en la probabilidad de supervivencia ($OR = 1.270, IC\ 95\ \%: 1.061 - 1.584, p = 0.016$), mientras que valores mas altos de PCR EGRE se relacionan con una disminución significativa en dicha probabilidad $OR = 0.970, IC\ 95\ \%: 0.947 - 0.987, p = 0.003$).

De igual forma, el valor de BUN mostró una asociación inversa con la supervivencia ($OR = 0.949, p = 0.047$), lo que indica que mayores niveles de urea en sangre podrían predecir un desbalance desfavorable. La PCR a las 24 horas también mostro un efecto positivo moderado ($OR = 1.019, p = 0.044$), lo que sugiere que su evolución en ese periodo puede ser indicativa de la respuesta del paciente al tratamiento.

Por otro lado, aunque variables como PCR ING ($p = 0.034$) también presentaron significancia, el resto de los predictores, como EDAD, GLUCOSA, CREATINA, etc., no mostraron asociaciones estadísticamente significativas ($p > 0.05$). Esto implica que bajo el modelo ajustado con los datos balanceados su influencia sobre el desbalance clínico no puede ser confirmada con los datos disponibles.

Tabla 6. Regresión logística (con balanceo SMOTE: *Oversampling*)

Características	β	Error estándar	Z	P	Exp (β)	OR
Intercepto	-1.12200	4.8322	-0.232	0.8160	0.326	(0.000–2904.635)
EDAD	-0.04291	0.0431	-0.994	0.3200	0.958	(0.686 – 1.004)
HTO	-0.04930	0.0505	-0.974	0.3290	0.952	(0.857 – 1.049
LEUCOS	0.23940	0.0992	2.414	0.0158	1.270	(1.061 – 1.584)

PLAQ	-0.00094	0.0028	-0.353	0.7230	0.999	(0.993 – 1.005)
GLUC	0.00113	0.0054	0.207	0.8360	1.001	(0.990 – 1.012)
BUN	-0.05236	0.0263	-1.987	0.0469	0.949	(0.895 – 0.996)
CREAT	0.62746	0.5470	1.147	0.2510	1.873	(0.651 – 5.971)
BILIS	0.36257	0.3595	1.009	0.3130	1.437	(0.739 – 3.147)
ALB	1.06922	0.9433	1.133	0.2570	2.913	(0.473 – 21.604)
COLT	-0.00517	0.0218	-0.236	0.8130	0.995	(0.948 – 1.036)
TRIG	0.00571	0.0086	0.663	0.5070	1.006	(0.898 – 1.024)
COL HDL	0.08178	0.0519	1.574	0.1160	1.085	(0.991 – 1.224)
COL LDL	-0.01568	0.0223	-0.703	0.4820	0.984	(0.940 – 1.028)
PCR ING	-0.02057	0.0096	-2.122	0.0339	0.980	(0.962 – 0.997)
PCR 24 HRS	0.01865	0.0092	2.015	0.0439	1.019	(1.001 – 1.039)
PCR EGRE	-0.03013	0.0101	-2.980	0.0029	0.970	(0.947 – 0.987)

Nota: OR = Odds ratio.

Los valores de p en rojo indican significancia estadística ($p < 0.01$); los valores en verde corresponden a asociaciones marginalmente significativas ($p < 0.05$).

Referencia: Elaboración propia

El valor del AIC obtenido en el modelo balanceado fue de 100.62, lo que indica un ajuste razonable del modelo con respecto a su complejidad. Aunque este valor es ligeramente superior al modelo sin balancear (AIC=84.75), el modelo balanceado logra mejorar la capacidad predictiva de la clase minoritaria al reducir el sesgo provocado por el desbalance de clases.

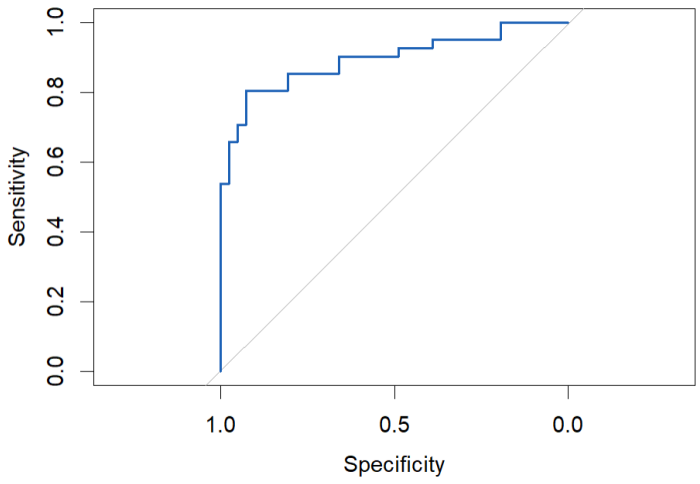

Figura 3. Área bajo la curva (sin balanceo)

Referencia: Elaboración propia.

La calidad del modelo se evalúa mediante el área bajo la curva ROC (AUC), cuyo valor fue de 0.8995, lo que refleja una alta capacidad discriminativa para distinguir entre pacientes que sobreviven y los que no. Esto se observa visualmente en la Figura 3 en donde la curva ROC se aleja de la línea diagonal que representa la clasificación aleatoria, indicando un buen desempeño del modelo.

Tabla 7. Matriz de confusión (con balanceo SMOTE: *Oversampling*)

Clase real	No sobrevivientes	Sobrevivientes
No sobrevivientes	37 (CC)	4 (FP)
Sobrevivientes	8 (FN)	33 (CC)

CC= Correctamente clasificados
FN= Falsos negativos
FP= Falsos positivos
Referencia: Elaboración propia

En cuanto a la capacidad predictiva, se obtuvo una precisión global del modelo (exactitud) de 85.4%, con una sensibilidad del 90.2 % (capacidad para identificar correctamente a los pacientes no sobrevivientes) y una especificidad del 90.2 % (capacidad para identificar a los sobrevivientes) de acuerdo con los valores presentados en la **Tabla 8**.

Tabla 8. Métricas por clase (con balanceo SMOTE: *Oversampling*)		
Métricas	Sobrevivientes	No sobrevivientes
Presión	80.5 %	82.2 %
Sensibilidad	80.5 %	90.2 %
Especificidad	90.2 %	80.5 %
F1-Puntaje	80.5 %	86.0 %
Área bajo de la curva ROC	0.895	0.895
Referencia: Elaboración propia		

4.2.2 Red bayesiana

En esta sección se presenta la aplicación de redes bayesianas para moderar la probabilidad de supervivencia en pacientes con sepsis. A continuación, se describen los resultados obtenidos con y sin el uso de técnica de balanceo de clase.

4.2.2.1 Red bayesiana sin balaceo

En primer momento, se construyó una red bayesiana empleando el conjunto de datos original, el cual presentaba un desbalance entre las clases. Este enfoque permite observar cómo la estructura probabilística se ajusta a los datos sin intervención, evaluando su capacidad predictiva y las relaciones aprendidas entre las variables clínicas bajo esta condición.

Para ello se realizaron los siguientes pasos para su análisis:

Parámetros de configuración del algoritmo HillClimber, en la **Figura 4** podemos observar los pasos.

- ❖ initAsNaiveBayes (Inicializar como Naive Bayes) → False
  - Si está en True, la red comienza como un modelo Naive Bayes (donde todas las variables dependen solo de la clase). Está en False, por lo que la red se aprende desde cero sin esta restricción.
- ❖ markovBlanketClassifier (Clasificador con Cobertura de Markov) → False
  - Si está en True, Weka intenta restringir la estructura de la red a la Cobertura de Markov, lo que reduce la complejidad. En este caso está en False, por lo que no se impone esta restricción.
- ❖ maxNrOfParents (Máximo número de padres) → 100000

- Es el límite máximo de nodos padres que puede tener cada variable en la red. Un valor de 100,000 significa que no se impone un límite práctico.

❖ scoreType (Métrica de evaluación) → BAYES

- Define el criterio para evaluar la calidad de una estructura de red. Aquí se usa Bayes, que evalúa la probabilidad del modelo dados los datos.

❖ useArcReversal (Permitir reversión de arcos) → False

- Si es True, el algoritmo puede invertir la dirección de los arcos en la red para mejorar el ajuste. En este caso está en False, lo que significa que una vez que un arco se establece, no puede cambiar de dirección.

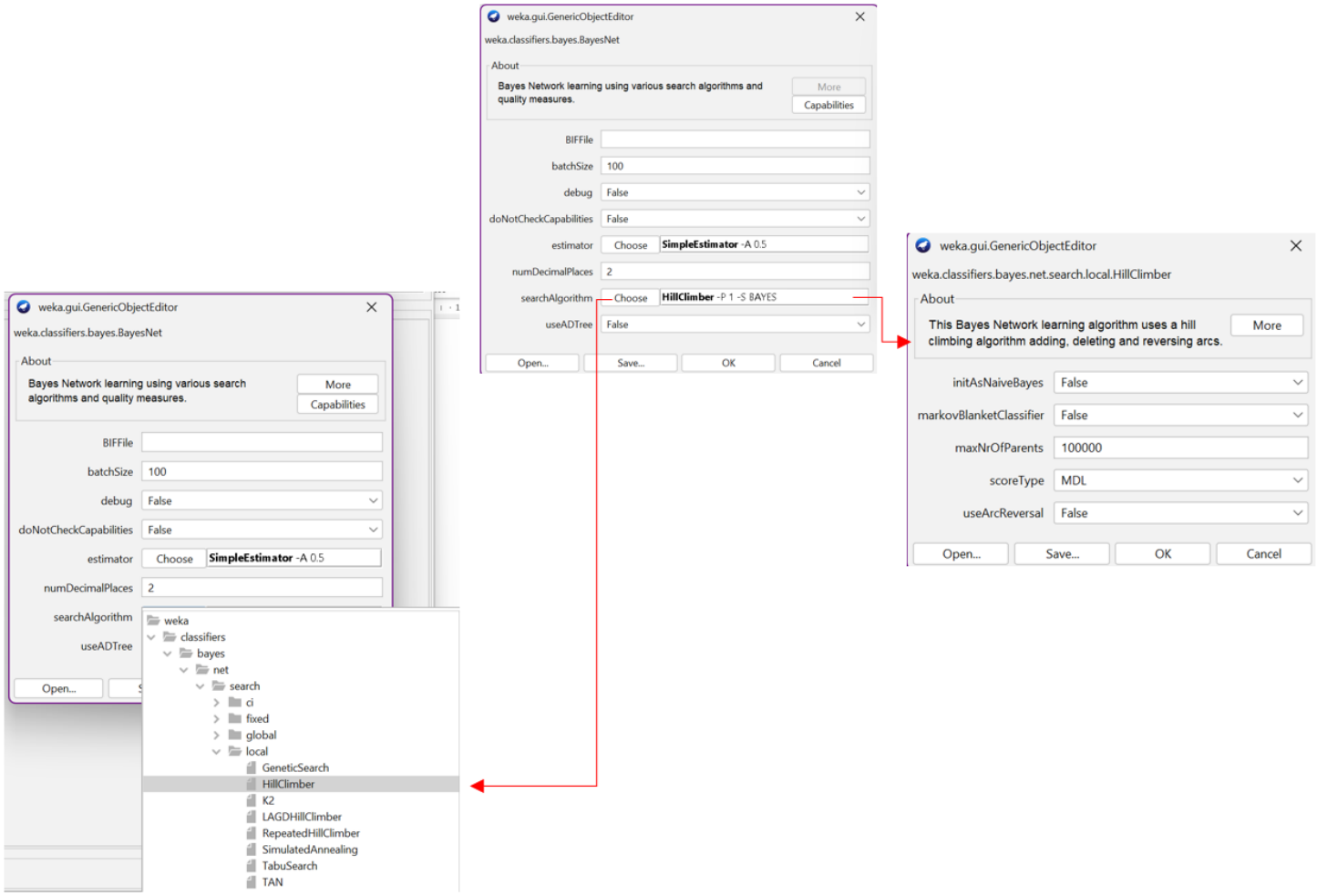

Figura 4. Pasos para definir los parámetros del algoritmo Hill Climber en Weka.

Referencia: Elaboración propia.

En la **Figura 5** podemos observar la red Bayesiana en donde la única variable directamente relacionada con el estado del paciente es la proteína C reactiva al egreso (PCR EGRE). Bajo este esquema, la probabilidad conjunta se puede descomponer como:

$$P(PCR_{EGRE}, ESTADO_{PAC}) = P(PCR_{EGRE} * P(ESTADO_{PAC}|PCR_{EGRE})$$

Figura 5. Grafo de la Red bayesiana creada (sin balanceo).

Referencia: Elaboración propia.

La precisión global del modelo clasifica **76.92 %** (50 de 65 instancias, están clasificadas correctamente).

La **Tabla 9** muestra la matriz de confusión obtenida donde el modelo logró una tasa de clasificación correcta del 76.9% con 33 verdaderos positivos (VP) para sobrevivientes y 17 verdaderos negativos (VN) para no sobrevivientes. No obstante, se identificaron 8 falsos positivos (FP) y 7 falsos negativos (FN), lo cual indica ciertas limitaciones del modelo al clasificar adecuadamente a ambas clases.

Tabla 9. Matriz de confusión (sin balanceo)

Predicción / Realidad	Sobrevivientes	No sobrevivientes
Sobrevivientes	33 (CC)	8 (FP)
No sobrevivientes	7 (FN)	17 (CC)

CC= Correctamente clasificados
FN= Falsos negativos
FP= Falsos positivos
Referencia: Elaboración propia

En la **Tabla 10**, se detallan las métricas por clase, se puede observar que el modelo alcanza una precisión del 82.5 % al predecir correctamente los pacientes sobrevivientes frente

a un 68 % para los no sobrevivientes. La sensibilidad, que se representa la capacidad de detectar correctamente los casos reales, es más alta para los sobrevivientes (80.5 %) que para los no sobrevivientes (70.8 %), lo que indica un mejor desempeño en la identificación de pacientes que superan la enfermedad.

Tabla 10. Métricas por clase (sin balanceo)		
Métricas	Sobrevivientes	No sobrevivientes
Presión	82.5 %	68.0 %
Sensibilidad	80.5 %	70.8 %
F1-Puntaje	81.5 %	69.4 %
Área bajo de la curva ROC	0.710	0.706
Referencia: Elaboración propia		

Finalmente, **la Tabla 11** presenta la distribución de probabilidades condicionales de la variable PCR EGRE. Se observa que cuando la proteína C reactiva al egreso es menor a 75.5, la probabilidad de supervivencia es del 85.2 %, mientas que esta probabilidad cae a 19.6 % cuando PCR EGRE supera dicho umbral. De manera inversa, la probabilidad de no supervivencia se incrementa drásticamente de 14.8 % a 80.4 % al pasar este umbral. Este patrón respalda el papel clave que desempeña la PCR al egreso como variable discriminante en el modelo.

Tabla 11. Probabilidades de la variable proteína C reactiva al egreso (sin balanceo)		
Estado del paciente	PCR EGRE < 75.5	PCR EGRE > 75.5
Sobrevivientes	0.852	0.196
No sobrevivientes	0.148	0.804
Referencia: Elaboración propia		

4.2.2.2 Red bayesiana con balanceo (SMOTE)

Para corregir esta discrepancia, se utilizó la técnica de "Remuestreo" un conjunto de datos aplicando la Técnica de Sobremuestreo de Minorías Sintéticas (SMOTE, por sus siglas en inglés). Con esta estrategia, se logró equiparar la cantidad de observaciones en ambas clases, permitiendo así una representación más equitativa de los datos durante el modelado.

Parámetros:

- ✓ classValue (Valor de la clase a balancear): Indica a qué clase se aplicará SMOTE. Si se deja en 0, el software detecta automáticamente la clase minoritaria.

- ✓ Debug (Modo de depuración): Si es True (Verdadero) mostrara información adicional sobre el proceso en la consola. En este caso está en False (Falso), por lo que no se imprimirá información extra.

- ✓ doNotCheckCapabilities (No verificar capacidades): Si es True, Weka omite la verificación de compatibilidad del filtro antes de ejecutarlo. Esto puede reducir el tiempo de ejecución, pero puede causar errores si los datos no son adecuados. En este caso está en False, lo que significa que Weka verificará primero si SMOTE se puede aplicar correctamente.

- ✓ nearestNeighbors (Vecinos más cercanos): Define cuántos vecinos más cercanos se usarán para generar cada nueva instancia sintética. En este caso, SMOTE tomará 5 vecinos para interpolar los nuevos datos sintéticos.

- ✓ Percentage (Porcentaje): Es el porcentaje de nuevas instancias sintéticas que se generarán. Se aplica sobre la cantidad original de la clase minoritaria. Con 24 observaciones originales, un 73% significa que se crearán aproximadamente 17 nuevas instancias, alcanzando un total de 41 instancias, igualando así la clase mayoritaria.

- ✓ randomSeed (Semilla aleatoria): Es el número utilizado para inicializar la generación de números aleatorios. Un mismo valor de semilla garantiza que los resultados sean reproducibles. Se puso 1, lo que significa que cada vez que se ejecute SMOTE con esta configuración, generará los mismos datos sintéticos, como se muestra en la **Figura 6**.

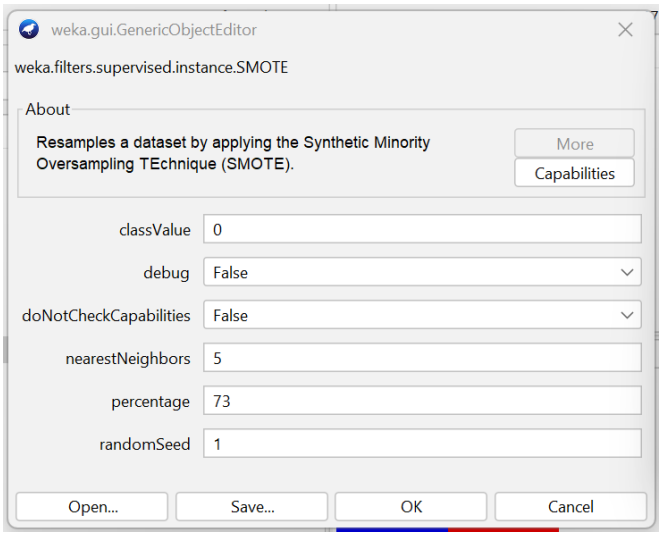

Figura 6. Parámetros para el balanceo de clases.

Referencia: Elaboración propia.

Aplicando estos parámetros se obtuvo adecuadamente el balanceo de las clases (**Figura 7**).

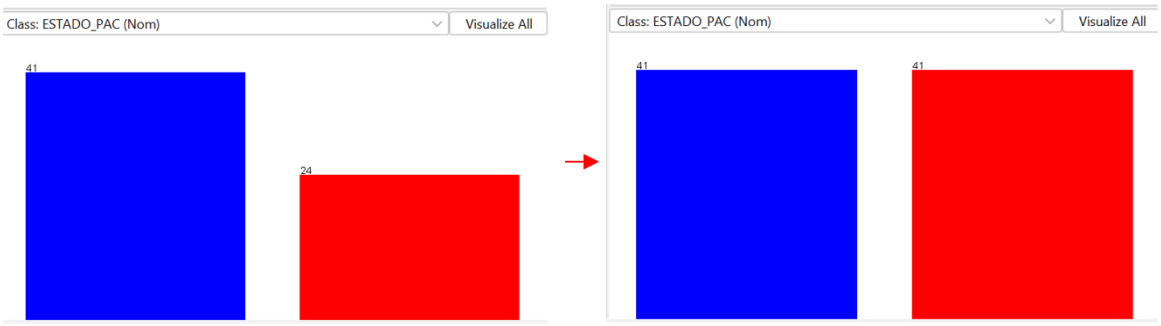

Figura 7. Datos sin balancear (imagen de la izquierda), datos balanceados (imagen de la derecha).

Referencia: Elaboración propia.

Una vez realizado el balanceo de las clases, se procedió a la construcción de la red bayesiana, con el procedimiento mencionado anteriormente con los datos sin balancear.

Ya con los parámetros definidos, procedemos a comenzar con los resultados. En la **Figura 8** podemos observar la red Bayesiana en donde las variables: colesterol total (COL T) y la proteína C reactiva al egreso (PCR EGRE) están directamente relacionados con el

estado del paciente (sobrevive o no). Bajo este esquema, la probabilidad conjunta se puede descomponer como:

$$P(COL_T, PCR_{EGRE}, COL_{HDL}, ESTADO_{PAC}) = P(COL_T) \cdot P(PCR_{EGRE} \mid COL_T) \cdot P(COL_{HDL} \mid PCR_{EGRE}) \cdot P(ESTADO_{PAC} \mid COL_T, PCR_{EGRE})$$

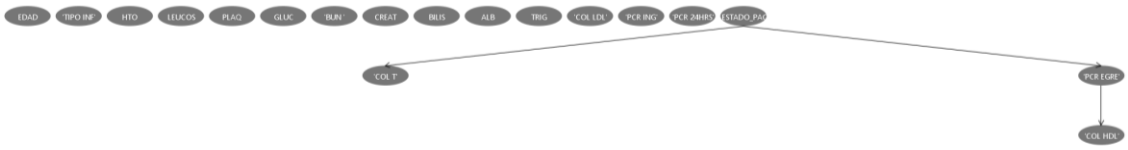

Figura 8. Grafo de la Red bayesiana creada (con balanceo).

Referencia: Elaboración propia.

La precisión global del modelo clasifica **80.49 %** (66 de 82 instancias están clasificadas correctamente).

La matriz de confusión la muestro en la **Tabla 12**.

Tabla 12. Matriz de confusión

Predicción / Realidad	Sobrevivientes	No sobrevivientes
Sobrevivientes	35 (CC)	6 (FP)
No sobrevivientes	10 (FN)	31 (CC)

CC= Correctamente clasificados
FN= Falsos negativos
FP= Falsos positivos
Referencia: Elaboración propia

Tabla 13. Métricas por clase

Métricas	Sobrevivientes	No sobrevivientes
Presión	77.8 %	83.8 %
Sensibilidad	85. 4 %	75.6 %
F1-Puntaje	81.4 %	79.5 %
Área bajo de la curva ROC	80. 9 %	80.9 %

Referencia: Elaboración propia

La precisión nos muestra las predicciones positivas (sobrevivientes o no), cuántas son correctas. Sensibilidad, cuántos de los sobrevivientes/no sobrevivientes fueron correctamente clasificados. F1-puntaje, balance entre la precisión y la sensibilidad. El área bajo la curva

muestra qué tan bien el modelo distingue entre ambas clases. Un 89.9 % es aceptable (**Tabla 13**).

Tabla 14. Probabilidades de la variable proteína C reactiva al egreso.		
Estado del paciente	PCR EGRE < 75.5	PCR EGRE > 75.5
Sobrevivientes	0.893	0.107
No sobrevivientes	0.202	0.798
Referencia: Elaboración propia		

La **Tabla 14** nos muestra las probabilidades de la variable **Proteína C Reactiva al Egreso** para los pacientes que sobrevivieron:

- Si **PCR EGRE** es menor a **75.5**, hay un **89.3% de probabilidad** de que el paciente haya sobrevivido.
- Si **PCR EGRE** es mayor a **75.5**, la probabilidad de supervivencia baja a **10.7%**.

Y para los pacientes que no sobrevivieron:

- Si **PCR EGRE** es menor a **75.5**, hay un **20.2% de probabilidad** de que no haya sobrevivido.
- Si **PCR EGRE** es mayor a **75.5**, la probabilidad de no sobrevivir sube a **79.8%**.

Una **PCR EGRE alta** está fuertemente asociada con un mayor riesgo de no supervivencia, mientras que una **PCR EGRE baja** es un buen predictor de supervivencia.

De la variable PCR EGRESO se desprende la variable COL HDL, esto quiere decir que la variable PCR EGRESO influye directamente sobre el COL HDL (Colesterol bueno), lo cual se representa mediante una relación condicional en la estructura de la red.

La **Tabla 15** muestra esta distribución condicional: cuando PCR EGRE es menor o igual a 75.5, la probabilidad de que COL HDL esté por encima de 31.53 es del 59.8 %. En cambio, cuando PCR EGRE es mayor a 75.5, dicha probabilidad se reduce significativamente a solo 11.8 %, mientras que la probabilidad de que COL HDL sea bajo (≤ 31.53) se incrementa a un 88.2 %.

Tabla 15. Probabilidades de supervivencia de la variable Colesterol bueno		
PCR EGRESO	COL HDL ≤ **31.53**	COL HDL > **31.53**
≤ **75**	0.402	0.598
> **75**	0.882	0.118
Referencia: Elaboración propia		

Tabla 16. Probabilidades de supervivencia de la variable Colesterol Total		
Estado del paciente	COL T < 127	COL T > 127
Sobrevivientes	0.44	0.56
No sobrevivientes	0.845	0.155
Referencia: Elaboración propia		

La **Tabla 16** nos muestra las probabilidades para la variable **Colesterol Total** para pacientes que sobrevivieron:

- Si **COL T** es menor a **127**, hay un **44% de probabilidad** de que el paciente haya sobrevivido.
- Si **COL T** es mayor a **127**, la probabilidad de supervivencia sube a **56%**.

Y para los pacientes que no sobrevivieron:

- Si **COL T** es menor a **127**, la probabilidad de no sobrevivir es **84.5%**.
- Si **COL T** es mayor a **127**, la probabilidad de no sobrevivir baja a **15.5%**.

Un **COL T bajo** (<127) está más asociado con la no supervivencia, mientras que un **COL T más alto** (>127) sugiere una mayor probabilidad de supervivencia.

Capítulo 5. Conclusiones

Discusión

Los resultados obtenidos en este estudio reflejan la viabilidad de aplicar técnicas estadísticas de aprendizaje automático en contextos clínicos con datos limitados. Se logró identificar variables clínicas asociadas significativamente al desenlace de pacientes con sepsis, destacando el valor de los modelos predictivos como herramientas complementarias para la toma de decisiones médicas.

La implementación del balanceo de clases mediante la técnica SMOTE permitió mitigar el sesgo presentado en los modelos, mejorando la sensibilidad y especificidad, especialmente en la regresión logística. Comparativamente, el modelo sin balanceo presentó una alta capacidad discriminativa general, pero tendía a favorecer a la clase mayoritaria. Con el balanceo, se alcanzó un rendimiento más equilibrado entre ambas clases.

En la regresión logística, variables como LEUCOS, BUN, PCR 24 HRS y PCR EGRE resultaron ser predictores significativos bajo el esquema balanceado, lo cual refuerza su importancia clínica. Por su parte, la red bayesiana mostró una mayor interpretabilidad del modelo: permitió visualizar relaciones condicionales relevantes, como la influencia del COL T sobre la PCR EGRE, y de esta última sobre el estado del paciente y variables asociadas como el COL HDL. Esta estructura evidenció que, aunque no establece causalidad determinista, conocer ciertos valores clínicos mejora considerablemente la predicción del desenlace, lo cual es útil para el diagnóstico clínico.

Además, el estudio enfrentó varias limitaciones. Primero, el tamaño reducido de la muestra (n=65) restringe la generalización de los hallazgos. Segundo, aunque se aplicaron técnicas de *oversampling*, estas generan datos sintéticos que podrían introducir un sesgo si no se usan adecuadamente. Por último, los modelos no fueron validados en contextos clínicos distintos, lo que representa una amenaza a la validez externa de los resultados. Como mecanismo de mitigación, se utilizó validación cruzada y técnicas robustas de ajuste y evaluación para mejorar la fiabilidad interna del análisis.

Conclusión

Esta investigación cumplió satisfactoriamente sus objetivos al desarrollar y validar modelos predictivos capaces de identificar factores clínicos relevantes en la detección temprana de sepsis, utilizando un conjunto limitado de datos reales del Instituto Nacional de Cardiología Dr. Ignacio Chávez. Se respondieron las preguntas de investigación al demostrar que tanto la regresión logística como las redes bayesianas son herramientas eficaces, con fortalezas complementarias: la primera en precisión y la segunda en interpretabilidad y modelado de relaciones condicionales.

El objetivo general se cumplió, ya que se desarrollaron dos modelos los cuales fueron la Regresión Logística y Redes Bayesianas con resultados buenos, incluso en condiciones de desbalance de clases y tamaño muestral reducido. La exactitud, sensibilidad y AUC obtenidas, validan su utilidad en contextos clínicos.

El primer objetivo particular se cumplió, ya que se identificaron aquellas variables significativas como LEUCOS, BUN, PCR 24 HRS y PCR EGRE. Estos predictores demostraron asociaciones clínicas consistentes con el desenlace del paciente, tanto en la regresión como en la red bayesiana. El siguiente objetivo nuevamente se cumplió, ya que se emplearon modelos adecuados para trabajar con una muestra pequeña de datos, y se implementó la técnica SMOTE para mitigar el desbalance. Y, por último, el tercer objetivo particular se cumple satisfactoriamente, ya que los modelos mostraron buenos niveles de desempeño. La regresión logística balanceada alcanzó una exactitud del 85.4 %, sensibilidad del 92. 2 % y AUC cercana al 0.90. La red bayesiana también logró más del 80 % de clasificación correcta tras el balanceo, confirmando su viabilidad.

Los modelos demostraron un buen rendimiento incluso en escenarios con limitaciones de datos, y su capacidad predictiva mejoró significativamente al aplicar balanceo de clases. Esto indica que es posible implementar soluciones analíticas avanzadas incluso en instituciones con recursos de datos limitados.

Cómo trabajos futuros se sugiere ampliar la muestra y aplicar validaciones externas en otros contextos clínicos, con el fin de mejorar la generalización y robustez de los modelos desarrollados. Asimismo, sería pertinente implementar técnicas de reducción de dimensionalidad, como el análisis de componentes principales, o la selección de variables basadas en criterios de información (AIC o BIC), para optimizar la parsimonia del modelo.

También se propone explorar modelos más complejos, como redes neuronales o bosques aleatorios, siempre cuidando la interpretabilidad. También se recomienda desarrollar herramientas visuales para implementar estos modelos con sistemas de alerta temprana en hospitales, fortaleciendo la toma de decisiones médicas basadas en datos.

Referencias

Banchón Alvarado, J. D., Saquicela, C. A. F., Nieto, J. M. V., & García, D. E. C. (2020). Conceptos actuales de sepsis y shock séptico. Journal of American health, 3(2), 102-116.

Baxter. (2025). Sepsis: enfermedad que afecta a más de 31 millones de personas en todo el mundo. Recuperado de https://www.baxter.mx/es/es/noticias-baxter/sepsis-enfermedad-que-afecta-mas-de-31-millones-de-personas-en-todo-el-mundo-y#_ftn2

Borges Sa, M. (2024). Detección precoz de sepsis (se) y shock séptico (ss) utilizando técnicas de big data, inteligencia artificial y machine learning.

Briceño, I. (2005). Sepsis: Definiciones y aspectos fisiopatológicos. Medicrit, 2(8), 164-178.

Cai, K., Lou, Y., Wang, Z., Yang, X., & Zhao, X. (2024). Machine Learning-Based Risk Prediction of Discharge Status for Sepsis. *Entropy*, *26*(8). https://doi.org/10.3390/e26080625

Chicco, D., y Jurman, G. (2020). Survival prediction of patients with sepsis from age, sex, and septic episode number alone. Scientific reports, 10(1), 17156.

Chitarroni, H. (2002). La regresión logística. Instituto de Investigación en Ciencias Sociales (IDICSO), Facultad de Ciencias Sociales, Universidad del Salvador.

Estupiñán, J. F. C., Junoy, F. N., y Diaz, J. M. O. (2016). Caracterización de pacientes diagnosticados con sepsis en una unidad de cuidados intensivos de Bucaramanga, Colombia 2010-2011: estudio descriptivo. Archivos de Medicina (Manizales), 16(1), 53-60.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian Network Classifiers. Machine Learning, 29(2–3), 131–163. https://doi.org/10.1023/A:1007465528199

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations, 11(1), 10–18.

Kosyakovsky, L. B., Somerset, E., Rogers, A. J., Sklar, M., Mayers, J. R., Toma, A., Szekely, Y., Soussi, S., Wang, B., Fan, C. P. S., Baron, R. M., & Lawler, P. R. (2022). Machine learning approaches to the human metabolome in sepsis identify metabolic links with

survival. *Intensive Care Medicine Experimental*, *10*(1). https://doi.org/10.1186/s40635-022-00445-8

López de Castilla-Vásquez, C. (2005). Clasificadores por redes bayesianas (Doctoral dissertation).

Toro Beltrán, C. F. (2022). Analítica a datos clínicos de pacientes de sepsis, estructurados bajo el estándar HL7 FHIR (CDA), facilitando la visualización en un dashboard para el diagnóstico oportuno.

Pale Carrion, R. A. (2006). "Caracteristicas clinicas y hemodinamicas de la sepsis en los pacientes cardiopatas". (Trabajo de grado de especialización). Universidad Nacional Autónoma de México, México. Recuperado de https://repositorio.unam.mx/contenidos/441374

Patel, S., Green, A., Wolfe, Y., Felock, G., Epstein, S., & Puri, N. (2023). The Impact of Positive Fluid Balance on Sepsis Subtypes: A Causal Inference Study. *Critical Care Research and Practice*, *2023*. https://doi.org/10.1155/2023/2081588

Rangel-Frausto, M. S. (1999). The epidemiology of bacterial sepsis. Infectious disease clinics of North America, 13(2), 299-312.

R Core Team. (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.r-project.org/

Ríos Bolaños, C. (2022). Diseño de un método de monitorización de pacientes con sepsis en Unidades de Cuidados Intensivos mediante técnicas de Machine Learning (Master's thesis).

Medicina, S. Y. (2019, 9 mayo). La mortalidad en sepsis aumenta un 8% por cada hora de retraso en aplicar el tratamiento. https://saludymedicina.org/post/la-mortalidad-en-sepsis-aumenta-un-8-por-cada-hora-de-retraso-en-aplicar-el-tratamiento

Sandoval Serrano, L. J. (2018). Algoritmos de aprendizaje automático para análisis y predicción de datos. Revista Tecnológica; no. 11.

Shaikhina, T., & Khovanova, N. A. (2017). Handling limited datasets with neural networks in medical applications: A small-data approach. *Artificial Intelligence in Medicine, 75*, 51–63. https://doi.org/10.1016/j.artmed.2016.12.003

Sierra Juárez, M. A., Quintana Barragán, K. P., Hernández Galván, J. A., Enríquez Sánchez, L. B., Pérez Ruiz, M. D., y Arzate Quintana, C. (2024). Validación de un modelo de inteligencia artificial para la predicción de la mortalidad del paciente con sepsis. Medicina Interna de México, 40(3).

Song, X., Liu, M., Waitman, L. R., Patel, A., y Simpson, S. Q. (2021). Clinical factors associated with rapid treatment of sepsis. Plos one, 16(5), e0250923.

Wang, J., Hu, Y., Zeng, J., Li, Q., He, L., Hao, W., Song, X., Yan, S., & Lv, C. (2023). Exploring the Causality Between Body Mass Index and Sepsis: A Two-Sample Mendelian Randomization Study. *International Journal of Public Health*, *68*. https://doi.org/10.3389/ijph.2023.1605548

WHO. (2024, 3 mayo). Word Health Organization. Sepsis. https://www.who.int/es/news-room/fact-sheets/detail/sepsis

“Lis de Veracruz: Arte, Ciencia, Luz”

www.uv.mx