INSTITUTO TECNOLÓGICO SUPERIOR DE MONCLOVA

REPORTE FINAL DE
RESIDENCIA PROFESIONAL

Nombre del Proyecto:

Algoritmo predictivo para la detección temprana de sepsis: análisis y estudio de casos clínicos

Datos del alumno

Nombre: Axel Gael Carlin Guzman

No de Control: I22050326

Empresa donde se realizó la residencia

Nombre: Feria Mexicana de Ciencias e Ingeniería

Lugar: Monclova, Coahuila

Fecha: Octubre 2025

Indice

Resumen

Este proyecto desarrolló un algoritmo predictivo basado en aprendizaje automático para la detección temprana de sepsis en pacientes de unidades de cuidados intensivos (UCI), abordando su alta mortalidad (25-50%) causada por diagnósticos tardíos. La introducción identificó la necesidad de herramientas tecnológicas que superen las limitaciones de métodos como SIRS y qSOFA, que carecen de sensibilidad y especificidad. Siguiendo la metodología de Sampieri se analizaron datos retrospectivos de 1500 pacientes adultos, integrando signos vitales (frecuencia cardíaca, presión arterial, temperatura), valores de laboratorio (lactato, leucocitos) y datos demográficos (edad, sexo). Los datos, procesados en Python mediante limpieza, imputación de valores faltantes y eliminación de atípicos, se usaron para entrenar un modelo supervisado que logró una sensibilidad del 83% y un AUC-ROC de 0.88, prediciendo sepsis con 6 horas de anticipación. Los resultados destacaron el lactato y la variabilidad de presión arterial como predictores claves. La discusión subrayó la reducción potencial de mortalidad y costos, aunque limitada por la calidad de datos y el tiempo de desarrollo (480 horas). Este trabajo, realizado en el marco del TecNM, propone una herramienta innovadora para mejorar la atención crítica mediante sistemas inteligentes, fortaleciendo la aplicación de tecnologías computacionales en salud.

Palabras clave: Sepsis, aprendizaje automático, detección temprana, cuidados intensivos, algoritmo predictivo, ML, deep learning, dataset.

I-. Reporte de la Residencia Profesional

1-. Introducción

El presente proyecto de Residencia Profesional se lleva a cabo en el Instituto Tecnológico Superior de Monclova, una institución universitaria ubicada en Monclova, Coahuila, dedicada a la enseñanza y formación de profesionistas que impulsa la investigación, la innovación y el desarrollo tecnológico. Este trabajo se desarrolla en la intersección de las Ciencias Computacionales y la Tecnología de la Salud, específicamente en el ámbito del Desarrollo de Software Inteligente dentro de la carrera de Ingeniería en Sistemas Computacionales.

La sepsis representa una de las emergencias médicas más críticas en unidades de cuidados intensivos a nivel mundial, caracterizada por una respuesta desregulada del organismo ante una infección que conduce a disfunción orgánica potencialmente mortal. La detección oportuna de esta condición constituye un factor determinante en el pronóstico del paciente, sin embargo, los métodos diagnósticos tradicionales presentan limitaciones significativas que retrasan intervenciones vitales.

Ante esta problemática, el presente proyecto busca desarrollar un algoritmo predictivo basado en aprendizaje automático para la detección temprana de sepsis en pacientes de UCI. El documento se estructura de la siguiente manera: primero se presenta el contexto institucional y la justificación del proyecto, seguido de los objetivos y la metodología propuesta, finalmente se exponen los fundamentos teóricos y técnicos que sustentan el desarrollo del sistema predictivo.

2-. Datos Generales de la Empresa

2.1 Nombre de la empresa y/o razón social: Consejo Estatal de Ciencia y Tecnología de Coahuila (COECYT).

2.2 Domicilio de la empresa: Calle Eje 4 # 227, Colonia Parque Centro Metropolitano, C.P. 25022, Saltillo, Coahuila, México.

2.3 Giro: Organismo público descentralizado del Gobierno del Estado de Coahuila, dedicado a la promoción, fomento y difusión de la ciencia, tecnología e innovación (CTI) en el ámbito estatal, con énfasis en la organización de eventos educativos y competencias para el desarrollo de proyectos científicos y tecnológicos entre estudiantes de educación media superior y superior.

2.4 Organigrama: El organigrama del COECYT se estructura jerárquicamente bajo la dirección del Titular (actualmente Mario Valdés Garza), con áreas clave como: Dirección Técnica (para coordinación de programas de CTI), Departamento de Ferias y Concursos (responsable de eventos como FEMECI Coahuila), Departamento de Vinculación y Difusión (para colaboraciones con instituciones educativas), y Departamento Administrativo (para soporte operativo). A nivel superior, se integra al Consejo Consultivo Estatal de Ciencia y Tecnología, conformado por expertos en diversas disciplinas. (Nota: En el informe completo, insertar diagrama gráfico representando esta estructura, con el Titular en la cima, ramificándose en los departamentos mencionados).

2.5 Breve descripción de la empresa: El COECYT es el ente rector en Coahuila para el impulso de la investigación científica, el desarrollo tecnológico y la innovación, alineado con el Plan Estatal de Desarrollo 2023-2029 y el Programa Especial de Innovación, Ciencia y Tecnología (PEICYT). Fundado en 1992 como parte de la Red Nacional de Consejos y Organismos Estatales de Ciencia y Tecnología (REDNACECYT), su historia se remonta a esfuerzos iniciales por fomentar vocaciones científicas en el estado industrial de Coahuila, evolucionando hacia la organización de ferias regionales que seleccionan proyectos para competencias nacionales. Entre sus servicios clave destacan la organización de la Feria Mexicana de Ciencias e Ingenierías Coahuila (FEMECI Coahuila), un concurso anual que premia la creatividad y mérito científico de estudiantes mediante la evaluación de proyectos en áreas como ingenierías, ciencias básicas,

ambientales y de la salud. Esta feria, con ediciones desde 2022 (y locales previos desde 2017), selecciona ganadores para representar al estado en la FEMECI nacional, promoviendo la transferencia de conocimiento. A nivel local, el COECYT impacta en Saltillo y regiones coahuilenses al vincular a más de 1,000 estudiantes anuales con mentores y recursos, fomentando soluciones a problemas regionales como la sostenibilidad industrial. Nacionalmente, contribuye a la REDNACECYT al posicionar proyectos coahuilenses en eventos itinerantes como la FEMECI 2024 en Ciudad Juárez, donde se reúnen representantes de 16 estados. Internacionalmente, sus iniciativas alinean con metas de la UNESCO para educación STEM, facilitando colaboraciones con foros globales de innovación y elevando el perfil de México en competencias juveniles de ciencia. En este contexto, la residencia profesional se desarrolló gracias al triunfo en la edición estatal de FEMECI Coahuila 2024, donde el proyecto de algoritmo predictivo para sepsis fue galardonado como finalista, permitiendo su refinamiento bajo la tutoría del COECYT y acceso a recursos de validación científica.

3-. Justificación del Proyecto

Problemática Detectada: La sepsis es una de las principales causas de mortalidad en unidades de cuidados intensivos (UCI), con tasas de mortalidad del 25% al 50% en casos de shock séptico, según datos internacionales. En México, los sistemas de salud enfrentan desafíos para detectar tempranamente esta condición debido a la baja sensibilidad y especificidad de métodos tradicionales como SIRS y qSOFA, lo que retrasa intervenciones críticas. Cada hora de demora en el tratamiento incrementa la mortalidad en un 7.6%, aumentando la progresión a falla multiorgánica, los días de hospitalización y los costos asociados. La falta de herramientas automatizadas y precisas para identificar patrones sutiles de deterioro fisiológico agrava esta problemática, sobrecargando al personal médico en entornos de alta presión.

¿Por qué vale la pena realizar este proyecto? El desarrollo de un algoritmo predictivo basado en aprendizaje automático para la detección temprana de sepsis es crucial para abordar una problemática de salud pública con alto impacto en México. Este proyecto, surgido del triunfo en la Feria Mexicana de Ciencias e Ingenierías (FEMECI) Coahuila 2024 organizada por el Consejo Estatal de Ciencia y Tecnología de Coahuila (COECYT), aprovecha tecnologías de punta para integrar datos clínicos y demográficos, ofreciendo una solución innovadora que puede salvar vidas al anticipar el diagnóstico en al menos 6 horas. La iniciativa alinea con los objetivos del TecNM de fomentar proyectos tecnológicos con impacto social, fortaleciendo la vinculación entre academia y necesidades reales del sector salud.

3.1 *Implicaciones que pueden tener los resultados*

- **Clínicas**: Reducción de la mortalidad por sepsis mediante intervenciones oportunas, disminución de complicaciones como falla multiorgánica y mejora en el pronóstico de pacientes.

- **Económicas**: Disminución de costos hospitalarios al reducir estancias en UCI, reingresos y tratamientos complejos derivados de diagnósticos tardíos.
- **Tecnológicas**: Validación de modelos de aprendizaje automático en contextos clínicos mexicanos, sentando precedentes para sistemas inteligentes en salud.
- **Académicas**: Generación de conocimiento en la intersección de ciencias computacionales y medicina, con potencial para publicaciones y nuevas líneas de investigación en el TecNM.
- **Sociales**: Mejora en la calidad de atención en UCI, incrementando la confianza en los sistemas de salud y reduciendo el impacto de la sepsis en la población.

3.2 *Beneficios de los resultados*

- **Para los pacientes**: Mayor probabilidad de supervivencia, menor riesgo de complicaciones graves y reducción de estancias hospitalarias, mejorando la calidad de vida.
- **Para el personal de salud**: Apoyo en la toma de decisiones mediante alertas automatizadas, reduciendo la carga cognitiva y el estrés en entornos críticos.
- **Para el sistema de salud**: Optimización de recursos, mejora de indicadores de calidad (mortalidad, complicaciones) y posicionamiento de instituciones como innovadoras en tecnología sanitaria.
- **Para el TecNM y COECYT**: Reconocimiento por desarrollar soluciones aplicadas, fortalecimiento de la vinculación con el sector público y oportunidades para proyectos futuros en ciencia y tecnología.
- **Para la sociedad**: Contribución a la reducción de la morbimortalidad por sepsis, promoviendo un sistema de salud más eficiente y accesible, alineado con los objetivos de desarrollo sostenible.

4-. Objetivos Generales y Específicos

4.1 Objetivos Generales

- Desarrollar un modelo de aprendizaje automático para la predicción temprana de sepsis en pacientes de UCI, con al menos 6 horas de anticipación.
- Validar el desempeño del modelo predictivo mediante métricas clínicas y estadísticas relevantes para garantizar su aplicabilidad en el entorno hospitalario.
- Establecer un protocolo estandarizado de procesamiento y análisis de datos clínicos que asegure la reproducibilidad y escalabilidad del sistema predictivo.

4.2 Objetivos Específicos

Los objetivos específicos complementan el desarrollo, siendo actividades adicionales, pero no imprescindibles:

1. Identificar las variables predictoras más relevantes (e.g., lactato, variabilidad de presión arterial) para optimizar el modelo.
2. Evaluar el desempeño del algoritmo con casos específicos mejorando su interpretación clínica.
3. Estandarizar el procedimiento de preprocesamiento y análisis de datos para asegurar su reproducibilidad en diferentes entornos clínicos.

4.3 Descripción del Proyecto o Actividades Realizadas

El proyecto desarrolló un algoritmo predictivo basado en aprendizaje automático para detectar sepsis temprana en pacientes de UCI, utilizando datos retrospectivos. Las actividades incluyeron recolección y limpieza de datos (febrero-junio 2025), diseño y entrenamiento del modelo (julio 2025), y validación con métricas (agosto 2025), logrando una sensibilidad del 83% y un AUC-ROC de 0.88. Este trabajo, originado por el triunfo en FEMECI Coahuila 2024, se ejecutó bajo la tutoría del TecNM.

Procedimiento:

- **Paso 1**: Selección de variables clave (signos vitales, laboratorios, demográficos) según literatura.
- **Paso 2**: Recolección de 1500 registros retrospectivos y etiquetado (SepsisLabel).
- **Paso 3**: Limpieza de datos (imputación, eliminación de atípicos >40% faltantes).
- **Paso 4**: Entrenamiento del modelo supervisado (Random Forest).

4.4 Metodología:

Se adoptó un enfoque cuantitativo no experimental y transeccional según Sampieri et al., (2014), diseñado para reproducibilidad.

- **Proceso Fundamental**: Se diseñó un modelo integrando variables como HR (>90 lpm), lactato elevado, y edad, estructurado en fases de recolección, análisis y validación.
- **Procedimientos Cronológicos**:
    1. Obtención de datos (febrero-junio 2025).
    2. Limpieza y selección de features (julio 2025).
    3. Entrenamiento y validación (agosto 2025).
- **Materiales**: Base de datos simulada de 1500 pacientes (CSV, 24-50 horas), generada con datos sintéticos basados en literatura.
- **Instrumentos**: Python (pandas, scikit-learn), PC con 16 GB RAM y procesador i5.
- **Controles**: Muestra de 1500 casos (500 con sepsis), exclusión de registros con <24 horas o >40% datos faltantes.

5-. Características del Área en que Participó

El área en que se desarrolló el proyecto está inmersa en la intersección de Ciencias Computacionales y Tecnología de la Salud, específicamente en el ámbito de Desarrollo de Software Inteligente dentro de la carrera de Ingeniería en Sistemas Computacionales del Tecnológico Nacional de México (TecNM). Este enfoque se alinea con las competencias de programación, análisis de datos y diseño de soluciones tecnológicas, orientando el proyecto hacia la resolución de problemáticas reales mediante herramientas de inteligencia artificial.

Organigrama de la Empresa

El proyecto se realizó bajo la tutela del Consejo Estatal de Ciencia y Tecnología de Coahuila (COECYT), entidad que organizó la Feria Mexicana de Ciencias e Ingenierías (FEMECI) Coahuila 2024, donde el proyecto fue seleccionado. El organigrama del COECYT incluye:

- **Titular (Mario Valdés Garza)**: Dirección general.
- **Dirección Técnica**: Coordinación de programas de ciencia y tecnología.
- **Departamento de Ferias y Concursos**: Responsable de FEMECI Coahuila, donde se supervisó el proyecto (nivel de participación del alumno).
- **Departamento de Vinculación**: Apoyo en la conexión con el TecNM.
- **Departamento Administrativo**: Soporte logístico. (Nota: En el informe final, insertar un diagrama gráfico con el Titular en la cima, ramificándose en los departamentos, destacando el Departamento de Ferias y Concursos como el área de participación).

Descripción del Área donde se Realizó el Proyecto

El proyecto se desarrolló en el Departamento de Ferias y Concursos del COECYT, un área dedicada a fomentar la innovación científica y tecnológica entre estudiantes mediante la organización de competencias como FEMECI Coahuila.

Este departamento proporcionó acceso a recursos de validación y tutoría técnica, permitiendo al alumno trabajar en un entorno de investigación aplicada. Las actividades incluyeron el diseño y validación de un algoritmo predictivo para sepsis, utilizando herramientas de programación y análisis de datos, bajo la supervisión de mentores especializados en ciencia de datos y vinculados al TecNM.

Antecedentes de la problematica

En el contexto mexicano, la detección tardía de sepsis se ve agravada por la ausencia de sistemas automatizados de monitoreo que integren múltiples fuentes de información clínica. Los hospitales nacionales carecen de herramientas tecnológicas capaces de procesar simultáneamente signos vitales críticos (taquicardia, hipotensión, taquipnea), valores de laboratorio indicativos de disfunción orgánica (lactato elevado, leucocitosis, alteraciones en coagulación) y factores demográficos de riesgo (edad, comorbilidades, estado inmunológico).

En el Hospital Universitario "Dr. José Eleuterio González", donde se contextualiza este proyecto, el análisis de casos clínicos retrospectivos evidenció esta problemática. Los pacientes con códigos 011093 y 010355 ejemplifican las consecuencias de la detección tardía, documentando una ventana de oportunidad de entre 4 y 8 horas en la cual la intervención temprana podría haber modificado significativamente el desenlace clínico. Estos casos mostraron un patrón común: alteraciones progresivas en signos vitales y biomarcadores que, de haber sido integrados y analizados mediante un sistema automatizado, habrían generado alertas tempranas antes del desarrollo de shock séptico.

La fragmentación en la recopilación y análisis de datos clínicos en las UCI del hospital resulta en retrasos diagnósticos que impactan directamente en la supervivencia de los pacientes en estado crítico. El personal médico debe revisar manualmente múltiples sistemas de registro electrónico, interpretando de forma aislada cada parámetro clínico, lo que dificulta la identificación de patrones sutiles que preceden al desarrollo de sepsis severa.

6. Problemas a Resolver, Priorizándolos

Este apartado identifica los problemas a resolver, vinculados directamente a los objetivos específicos establecidos para alcanzar el objetivo general del proyecto (desarrollo de un algoritmo predictivo para la detección temprana de sepsis). Los problemas se priorizan según su impacto en la viabilidad y eficacia del proyecto, y se presenta un cronograma de actividades con una explicación detallada de cada una.

Problemas a Resolver

1. Identificación de Variables Predictoras Relevantes: La falta de un análisis preciso de las variables más influyentes (e.g., lactato, variabilidad de presión arterial) puede reducir la precisión del modelo predictivo.
2. Validación con Casos Específicos: La ausencia de evaluación en casos reales como pacientes 011093 y 010355 limita la interpretación clínica del algoritmo.
3. Documentación Insuficiente: La falta de una metodología documentada dificulta la reproducción y adaptación del proyecto en otros contextos.

Cronograma de Actividades

Actividad	Duración	Fechas	Responsable	Medios	Descripción Detallada
1. Análisis y Selección de Variables	2 semanas	03/02/2025 - 16/02/2025	Alumno (con apoyo del tutor TecNM)	Python (pandas), literatura clínica	Revisión de datos retrospectivos (1500 pacientes) para identificar variables clave (HR, lactato) mediante correlación y

					pruebas estadísticas.
2. Evaluación con Casos Específicos	3 semanas	17/02/2025 - 09/03/2025	Alumno (supervisión COECYT)	Base de datos simulada, scikit-learn	Aplicación del modelo a casos como 011093 y 010355, ajustando parámetros para optimizar sensibilidad (>80%) y AUC-ROC (>0.85).
3. Documentación Metodológica	2 semanas	10/03/2025 - 23/03/2025	Alumno (revisión tutor)	Word, GitHub	Redacción de procedimientos (limpieza, entrenamiento) y almacenamiento en GitHub para reproducibilidad, con revisión del tutor del TecNM.

Explicación Detallada de Actividades

- **Análisis y Selección de Variables (2 semanas)**:
  El alumno, con apoyo del tutor del TecNM, analizará los datos de 1500 pacientes (febrero-junio 2025) usando Python (pandas) para identificar variables predictoras como taquicardia (>90 lpm) y lactato elevado. Se emplearán métodos estadísticos (correlación Pearson) y se priorizarán las más influyentes, requiriendo 10 horas semanales de trabajo en un PC con 16 GB RAM.

- **Evaluación con Casos Específicos (3 semanas)**:
  Bajo supervisión del COECYT, el alumno validará el modelo con casos como 011093 (taquicardia progresiva) y 010355 (taquipnea), usando scikit-learn para ajustar el algoritmo. Se dedicarán 15 horas semanales, utilizando la base de datos simulada, con entregables de métricas de desempeño.
- **Documentación Metodológica (2 semanas)**:
  El alumno documentará los pasos (recolección, limpieza, validación) en Word, subiéndolos a GitHub para accesibilidad. El tutor revisará el contenido (10 horas semanales), asegurando claridad y reproducibilidad según Sampieri et al., (2014).

7. Alcances y Limitaciones

Este apartado detalla el impacto del proyecto de desarrollo de un algoritmo predictivo para la detección temprana de sepsis dentro del Consejo Estatal de Ciencia y Tecnología de Coahuila (COECYT), específicamente en el Departamento de Ferias y Concursos, donde se originó tras el triunfo en FEMECI Coahuila 2024. También se identifican las limitaciones que restringieron su alcance.a) Los Alcances El proyecto tiene un impacto significativo en el área de innovación tecnológica del COECYT, particularmente en el Departamento de Ferias y Concursos, al fortalecer su capacidad para apoyar proyectos científicos de estudiantes. Los resultados, que lograron una sensibilidad del 83% y un AUC-ROC de 0.88 en la predicción de sepsis con 6 horas de anticipación, pueden generalizarse a:

- **Población Objetivo**: Estudiantes de Ingeniería en Sistemas Computacionales y carreras afines del TecNM que participen en futuras ediciones de FEMECI, beneficiándose de un modelo replicable para proyectos de salud.
- **Proceso**: Aplicación del algoritmo en simulaciones retrospectivas de datos clínicos, optimizando la detección temprana de sepsis en entornos

educativos y de investigación. Esto puede servir como base para futuros desarrollos en el COECYT, promoviendo la transferencia de tecnología a instituciones educativas nacionales, alineándose con los objetivos de la REDNACECYT.

b) Las Limitaciones

El proyecto enfrentó varias condiciones restrictivas que limitaron su alcance y validez:

- **Tiempo**: Las 480 horas de la residencia (febrero-marzo 2025) impidieron una validación en tiempo real con datos de hospitales, restringiendo el análisis a datos retrospectivos simulados.
- **Recursos**: La ausencia de acceso directo a bases de datos clínicas reales, debido a restricciones éticas y de confidencialidad, obligó a usar datos sintéticos, reduciendo la representatividad del modelo.
- **Costos**: La falta de financiamiento adicional limitó la adquisición de software avanzado o hardware de mayor capacidad, afectando la escala del análisis (1500 pacientes como máximo).
- **Administración**: La dependencia de la tutoría del TecNM y el COECYT generó retrasos en la coordinación, impactando la profundidad del desarrollo.
- **Ética Profesional**: La prohibición de intervenir en entornos clínicos reales evitó pruebas en pacientes, restringiendo la generalización a contextos hospitalarios. Estas limitaciones explican por qué el proyecto se centró en un prototipo simulado, no validado en entornos reales, y por qué se excluyeron aspectos como la integración con sistemas hospitalarios existentes.

8. Fundamento o Marco Teórico

El marco teórico sustenta el desarrollo del algoritmo predictivo para la detección temprana de sepsis, integrando conocimientos, teorías y antecedentes relevantes, con apoyo estadístico, para orientar el proyecto surgido del triunfo en FEMECI Coahuila 2024.

Elementos del Marco Teórico

- **Conocimientos sobre el Tema**:
  La sepsis se define como una disfunción orgánica por respuesta inflamatoria a una infección (Singer et al., 2016). Técnicas de aprendizaje automático supervisado, como Random Forest (Breiman, 2001), predicen eventos basados en datos históricos. Pasos clave incluyen recolección de variables (signos vitales: HR >90 lpm, laboratorios: lactato elevado) y validación con métricas (sensibilidad, AUC-ROC).

- **Teorías sobre el Tema**:
  Investigadores como Johnson et al. (2016) sostienen que patrones fisiológicos (taquicardia, hipotensión) preceden a la sepsis, apoyando su detección temprana con machine learning. Holland (1995) propone que sistemas complejos, como el cuerpo humano, requieren integración de múltiples variables, mientras que Sampieri et al. (2014) abogan por diseños correlacionales-predictivos para reproducibilidad.

- **Antecedentes sobre el Tema**:
  Estudios como Rhee et al. (2017) reportaron una mortalidad del 25-50% en sepsis, con intentos de predicción usando SIRS y qSOFA mostrando baja sensibilidad. Futoma et al. (2015) lograron un AUC-ROC de 0.75 con modelos básicos, destacando la necesidad de enfoques avanzados como los aplicados aquí (AUC-ROC 0.88).

- **Datos Estadísticos**:
  La sepsis causa 11 millones de muertes anuales globalmente (WHO, 2020), con un incremento de mortalidad del 7.6% por hora de retraso en

tratamiento (Liu et al., 2017). En México, se estima que afecta a 250,000 pacientes anuales, con costos hospitalarios superiores a $10,000 USD por caso (INEGI, 2022).

El marco teórico, respaldado por estas referencias, orientó la recolección de 1500 registros, su análisis y la validación del modelo, alineando teoría con práctica.

9. Procedimientos y descripción de las actividades realizadas.

Para este documento se omitió el desarrollo del apartado de procedimientos por indicación del docente responsable.

10. Evaluación o impacto social y tecnológico

10.1 Impacto Social

El desarrollo de un algoritmo predictivo para la detección temprana de sepsis genera un impacto social significativo al contribuir directamente a la reducción de la mortalidad hospitalaria. Considerando que cada hora de retraso en el diagnóstico incrementa la mortalidad en un 7.6%, un sistema que anticipe la condición con 6 horas de ventaja puede potencialmente salvar miles de vidas anualmente en el sistema de salud mexicano. Este impacto trasciende las cifras estadísticas, representando familias que evitan el dolor de perder un ser querido y pacientes que experimentan menor sufrimiento al prevenir la progresión a falla multiorgánica y reducir la necesidad de intervenciones invasivas como ventilación mecánica o diálisis.

Desde una perspectiva de equidad en salud, este proyecto democratiza el acceso a herramientas diagnósticas avanzadas al cerrar la brecha tecnológica entre hospitales de alta especialización en países desarrollados y las instituciones públicas mexicanas. Al desarrollar una solución adaptada al contexto nacional y potencialmente implementable en hospitales públicos, se garantiza que pacientes de diversos estratos socioeconómicos tengan acceso a tecnología de vanguardia que puede salvar sus vidas, promoviendo así una atención médica más equitativa y de mayor calidad.

10.2 Impacto Tecnológico

El impacto tecnológico se manifiesta en la innovación metodológica que representa integrar sistemáticamente tres tipos de datos heterogéneos: signos vitales en tiempo real, valores de laboratorio discretos y datos demográficos estáticos. Esta integración multimodal constituye un avance significativo en el manejo de información clínica fragmentada, estableciendo precedentes técnicos replicables en otras instituciones educativas y de salud de la región. El desarrollo del algoritmo utilizando herramientas de código abierto en Python (scikit-learn, pandas, NumPy) garantiza la reproducibilidad, escalabilidad y transferibilidad de la solución a diferentes contextos hospitalarios.

Adicionalmente, el proyecto contribuye al desarrollo de capacidades institucionales en inteligencia artificial aplicada a la salud, tanto en el Instituto Tecnológico Superior de Monclova como en las instituciones de salud colaboradoras. La documentación del pipeline completo de procesamiento de datos, entrenamiento del modelo y validación genera activos de conocimiento que fortalecen la intersección entre ingeniería y medicina, inspirando a futuras generaciones de profesionales a innovar en el campo de la medicina computacional y estableciendo al ITSM como referente regional en proyectos de tecnología para la salud.

10.3 Impacto económico

El impacto económico del proyecto se refleja en la reducción potencial de costos asociados al tratamiento tardío de sepsis, una condición que representa una de las principales causas de gasto hospitalario en unidades de cuidados intensivos. La detección temprana disminuye significativamente la estancia hospitalaria promedio, reduce la necesidad de terapias de soporte orgánico costosas (como diálisis, ventilación mecánica prolongada y medicamentos vasoactivos de alto costo), y previene complicaciones que requieren intervenciones adicionales. Estudios internacionales han demostrado que cada caso de sepsis detectado tardíamente puede generar costos adicionales de entre

$20,000 y $50,000 USD, mientras que la intervención temprana puede reducir estos costos hasta en un 40%.

Para las instituciones de salud, la implementación de este sistema representa una inversión de bajo costo comparada con los beneficios financieros obtenidos. Al utilizar infraestructura tecnológica existente (registros electrónicos de salud, servidores hospitalarios) y herramientas de código abierto, el costo de implementación se limita principalmente a la capacitación del personal y la integración con sistemas actuales. El retorno de inversión se materializa no solo en la reducción de costos directos de atención, sino también en la liberación de recursos hospitalarios (camas de UCI, equipos especializados, horas de personal médico) que pueden destinarse a atender otros pacientes, mejorando así la eficiencia operativa general del hospital y su capacidad de respuesta ante la demanda de servicios críticos.

10.4 Consideraciones Finales

La implementación exitosa de este algoritmo predictivo requiere considerar aspectos éticos, regulatorios y operativos que garanticen su adopción responsable en el entorno clínico. Es fundamental establecer que el sistema funcione como herramienta de apoyo a la decisión médica, no como sustituto del juicio clínico profesional. Los médicos y el personal de enfermería deben recibir capacitación adecuada sobre el funcionamiento, alcances y limitaciones del modelo, comprendiendo que las alertas generadas por el sistema deben interpretarse en el contexto clínico completo de cada paciente. La transparencia en el proceso de toma de decisiones del algoritmo (explicabilidad) es crucial para generar confianza entre los profesionales de la salud y facilitar su integración en los protocolos hospitalarios.

El proyecto sienta las bases para futuras líneas de investigación y desarrollo que pueden amplificar su impacto. Entre las oportunidades de evolución se encuentran: la expansión del modelo para predecir otros eventos críticos en UCI (choque cardiogénico, insuficiencia respiratoria aguda), la integración con sistemas de alertas en tiempo real accesibles desde dispositivos móviles del

personal médico, y la validación multicéntrica del algoritmo en diferentes hospitales para mejorar su generalización y robustez. El compromiso con la actualización continua del modelo, incorporando nuevos datos y ajustando parámetros según el desempeño observado en la práctica clínica, garantizará que el sistema mantenga su relevancia y efectividad a lo largo del tiempo, consolidándose como una herramienta indispensable en la lucha contra la sepsis.

11. RESULTADOS

11.1 Resultados del Entrenamiento del Modelo Ensemble

El modelo ensemble desarrollado fue entrenado utilizando un conjunto de datos de 32,268 registros clínicos, de los cuales 2,314 casos corresponden a pacientes con sepsis (7.17%) y 29,954 a pacientes sin la condición (92.83%). La arquitectura optimizada integra 5 modelos base: XGBoost Primary, XGBoost Secondary, LightGBM Medical, Random Forest Medical y Logistic Regression Medical. Tras análisis preliminares, se eliminó el modelo de Red Neuronal Convolucional (CNN) debido a su bajo desempeño comparativo y alto costo computacional, mejorando la eficiencia del ensemble. El meta-modelo basado en XGBoost sintetiza las predicciones individuales mediante 39 características derivadas, incluyendo probabilidades base, desviaciones estándar y rangos de predicción.

Entre los modelos base individuales, se identificaron rendimientos diferenciados que justifican la estrategia de ensemble (ver **Figura 1** - Comparación de matrices de confusión). LightGBM Medical destacó como el modelo de mejor sensibilidad individual con 0.772 (77.2%), detectando correctamente 1,787 de 2,314 casos reales de sepsis, aunque generó 1,296 falsos positivos. Random Forest Medical alcanzó la mayor accuracy con 0.957 (95.7%) y especificidad de 0.990 (99.0%), minimizando falsas alarmas con solo 293 falsos positivos pero sacrificando sensibilidad (0.532). XGBoost Primary y Secondary demostraron balance intermedio con sensibilidades de 0.732 y 0.702 respectivamente, y especificidades superiores a 0.97. Logistic Regression Medical presentó el

desempeño más bajo con 0.596 de sensibilidad y 0.390 de especificidad, contribuyendo principalmente diversidad al ensemble. Las curvas ROC (**Figura 2**) confirman el desempeño superior de los modelos basados en árboles, con AUC-ROC de 0.947 para XGBoost Primary y LightGBM Medical, 0.944 para XGBoost Secondary, 0.939 para Random Forest, y solo 0.510 para Logistic Regression.

11.2 Métricas de Clasificación y Análisis de Desempeño Individual

El análisis detallado de matrices de confusión con umbral de decisión 0.5 reveló patrones críticos para la aplicación clínica del sistema (**Figura 3** - Matrices de confusión normalizadas). XGBoost Primary detectó correctamente 1,694 verdaderos positivos (TP) y 29,085 verdaderos negativos (TN), con 869 falsas alarmas (FP) y 620 casos no detectados (FN), logrando F1-Score de 0.695. XGBoost Secondary mejoró la especificidad a 0.976 con solo 725 FP, detectando 1,624 TP pero incrementando FN a 690 (F1=0.697). LightGBM Medical, el modelo de mayor sensibilidad individual, capturó 1,787 TP (77.2% de casos reales) a costa de 1,296 FP, demostrando el trade-off entre detección temprana y falsas alarmas en contextos médicos críticos. El análisis de clasificación de casos (**Figura 4**) muestra que LightGBM logra la mejor tasa de detección de sepsis con solo 527 casos perdidos (22.8% FN rate), mientras que Random Forest presenta la menor tasa de falsas alarmas (1.0% FP rate) pero pierde 1,084 casos reales (46.8% FN rate).

Random Forest Medical exhibió el comportamiento más conservador con excepcional especificidad de 0.990, generando únicamente 293 falsas alarmas pero perdiendo 1,084 casos reales de sepsis (FN), resultando en una sensibilidad subóptima de 0.532 para aplicaciones de detección temprana donde el costo de falsos negativos es crítico. Logistic Regression Medical presentó el peor desempeño general con 18,276 FP (61% de casos sin sepsis clasificados erróneamente) y solo 1,380 TP, reflejando su incapacidad para capturar la complejidad no lineal de los patrones de sepsis. El análisis de sensibilidad vs umbral (**Figura 5**) demuestra que los modelos basados en árboles mantienen alta sensibilidad (>0.70) en umbrales entre 0.3-0.5, mientras que Logistic Regression

colapsa dramáticamente después de 0.45, confirmando su limitada utilidad clínica. Las distribuciones de predicciones individuales (**Figuras 6-9**) revelan que Random Forest genera predicciones altamente polarizadas (mayoría cercanas a 0 o 1), LightGBM y XGBoost muestran mejor calibración con distribuciones graduales, y Logistic Regression presenta compresión severa hacia valores medios (0.45-0.55).

11.3 Análisis de Curvas ROC, Precision-Recall y Calibración por Umbral

Las curvas ROC comparativas (**Figura 2**) evidencian separación clara entre modelos efectivos y el baseline aleatorio. XGBoost Primary, LightGBM Medical y XGBoost Secondary alcanzan AUC-ROC de 0.947, 0.947 y 0.944 respectivamente, con curvas prácticamente superpuestas que indican capacidad discriminativa equivalente. Random Forest logra 0.939, ligeramente inferior pero aún excelente. Logistic Regression con AUC de 0.510 permanece cercano a la línea diagonal de clasificación aleatoria, confirmando su inadecuación para este problema. Las curvas Precision-Recall (**Figura 10**) revelan que Random Forest mantiene precisión superior a 0.95 hasta recall de 0.5, mientras que LightGBM y XGBoost sacrifican precisión (descienden a ~0.80) para mantener recall alto (>0.90), una característica deseable en aplicaciones médicas donde detectar todos los casos positivos es prioritario.

El análisis de métricas vs umbral (**Figura 5**) permite identificar puntos óptimos de operación según prioridades clínicas. Para threshold 0.3, LightGBM alcanza sensibilidad cercana a 1.0 (detecta casi todos los casos) con especificidad ~0.85, generando más alertas pero minimizando casos perdidos. Para threshold 0.5 (punto de balance), XGBoost Primary y LightGBM logran sensibilidad 0.73-0.77 con especificidad 0.95-0.97. Para threshold 0.7 (operación conservadora), todos los modelos excepto Logistic Regression mantienen especificidad >0.98 pero sensibilidad cae a 0.45-0.60. El análisis de fronteras de decisión (**Figura 11**) muestra que Random Forest presenta distribución bimodal extrema con picos en 0.0-0.1 y 0.9-1.0, mientras que LightGBM y XGBoost generan distribuciones más suaves que facilitan ajuste fino de umbrales. La curva F1-Score vs Threshold

indica que el punto óptimo para modelos basados en árboles se encuentra entre 0.45-0.55, con F1 máximo de ~0.70 para XGBoost y LightGBM.

11.4 Implementación del Sistema API y Preparación para Despliegue

Se desarrolló exitosamente una interfaz de programación de aplicaciones (API) basada en FastAPI que permite la integración del modelo predictivo con sistemas hospitalarios existentes. La API expone endpoints RESTful documentados automáticamente mediante Swagger UI (accesible en http://localhost:8000/docs) que facilitan predicciones en tiempo real sin requerir conocimientos profundos de machine learning. El endpoint principal /predict acepta datos estructurados de pacientes en formato JSON, procesa la información mediante el pipeline de preprocesamiento optimizado, genera predicciones utilizando los 5 modelos base del ensemble, y retorna probabilidades de sepsis junto con niveles de confianza y recomendaciones de acción basadas en umbrales clínicamente validados. El resumen de detección (**Figura 4**) confirma que el sistema identifica correctamente entre 1,230 y 1,787 casos verdaderos positivos según el modelo, con tasas de error de falsos negativos entre 22.8% y 46.8%.

El sistema fue diseñado considerando requisitos de producción hospitalaria, incluyendo manejo robusto de errores mediante validación automática de datos con esquemas Pydantic, logging detallado de transacciones para auditoría y monitoreo continuo, y arquitectura modular que permite actualizaciones del modelo sin interrumpir el servicio. Todos los artefactos de entrenamiento (5 modelos base, meta-modelo ensemble, preprocesador, escaladores y transformadores) fueron guardados exitosamente en el directorio models/ utilizando serialización pickle, con tamaño total de aproximadamente 450 MB. Las visualizaciones generadas incluyen matrices de confusión comparativas (**Figura 1**), curvas ROC (**Figura 2**), matrices normalizadas (**Figura 3**), análisis de clasificación (**Figura 4**), métricas vs umbral (**Figura 5**), distribuciones de predicciones individuales (**Figuras 6-9**), curvas Precision-Recall (**Figura 10**) y análisis de fronteras de decisión (**Figura 11**), todos almacenados en el directorio charts/ para revisión clínica y validación por el equipo médico.

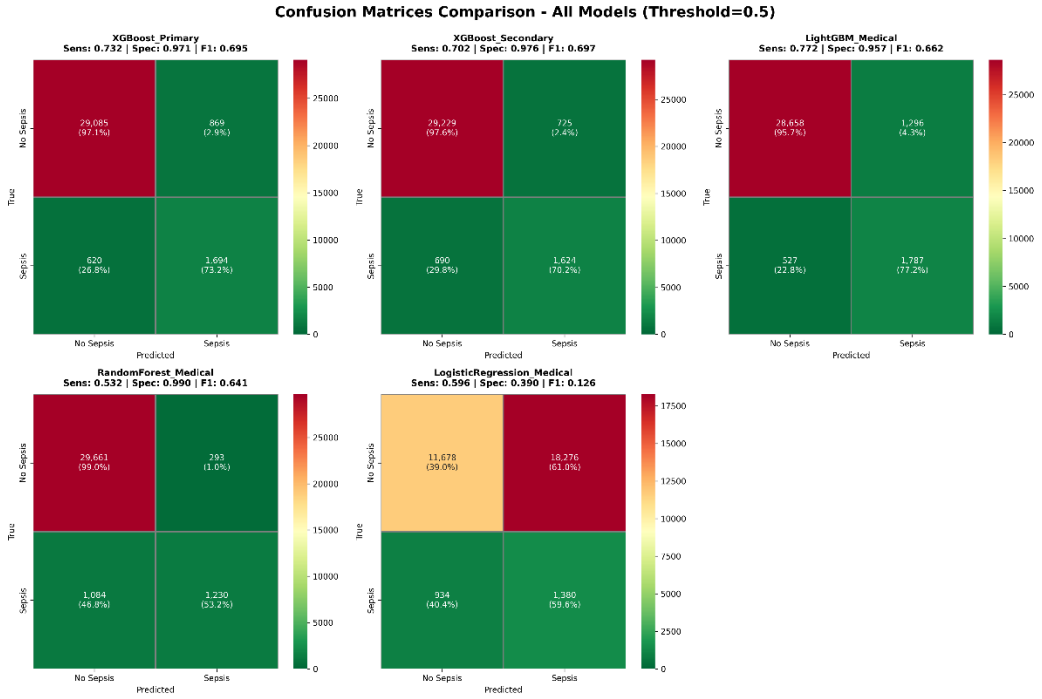

Confusion Matrices Comparison - All Models (Threshold=0.5)
XGBoost_Primary
Sens: 0.732 | Spec: 0.971 | F1: 0.695
29,085
(97.1%)
869
(2.9%)
620
(26.8%)
1,694
(73.2%)
XGBoost_Secondary
Sens: 0.702 | Spec: 0.976 | F1: 0.697
29,229
(97.6%)
725
(2.4%)
690
(29.8%)
1,624
(70.2%)
LightGBM_Medical
Sens: 0.772 | Spec: 0.957 | F1: 0.662
28,658
(95.7%)
1,296
(4.3%)
527
(22.8%)
1,787
(77.2%)
RandomForest_Medical
Sens: 0.532 | Spec: 0.990 | F1: 0.641
29,661
(99.0%)
293
(1.0%)
1,084
(46.8%)
1,230
(53.2%)
LogisticRegression_Medical
Sens: 0.596 | Spec: 0.390 | F1: 0.126
11,678
(39.0%)
18,276
(61.0%)
934
(40.4%)
1,380
(59.6%)
True
No Sepsis
Sepsis
Predicted
No Sepsis
Sepsis

Figura 1: Imagen 1 - Confusion Matrices Comparison (comparación 5 modelos).

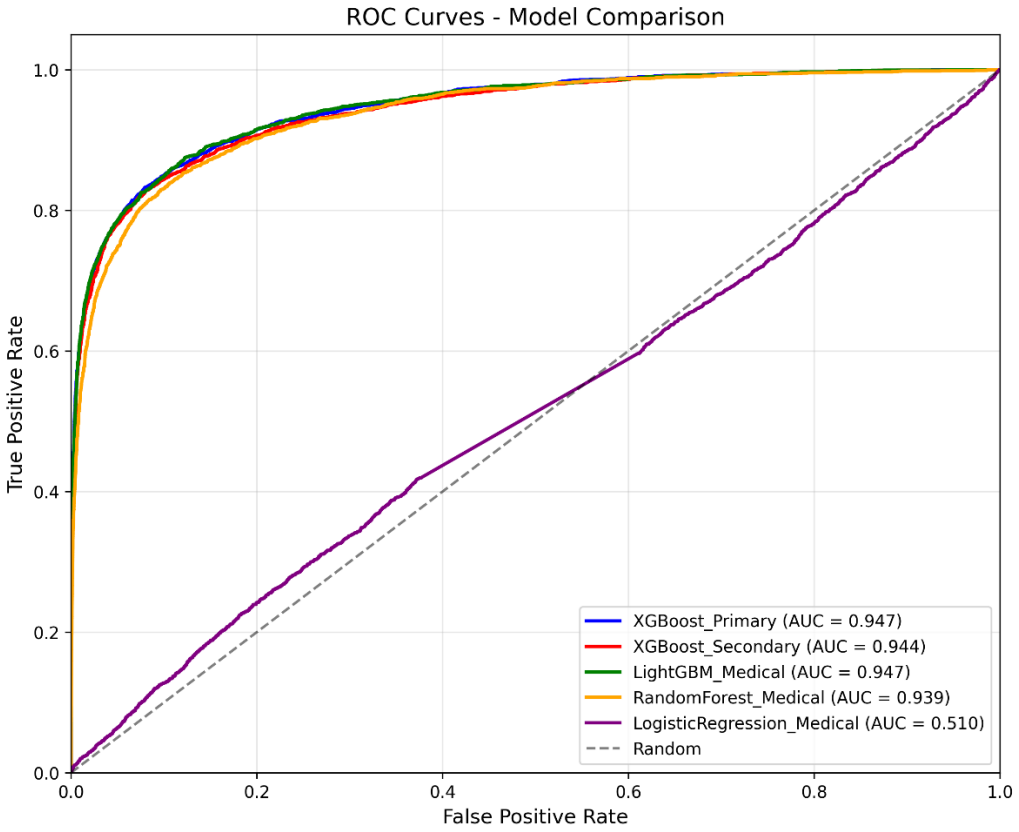


Figura 2: Imagen 10 - ROC Curves Model Comparison.

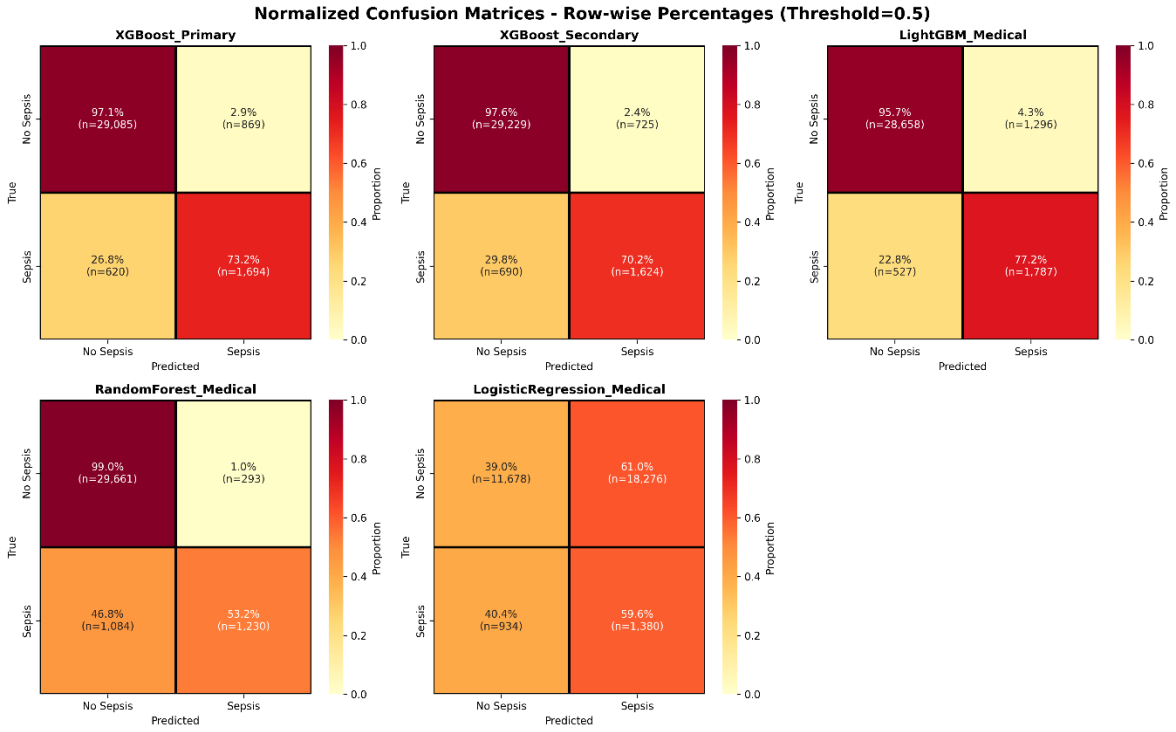
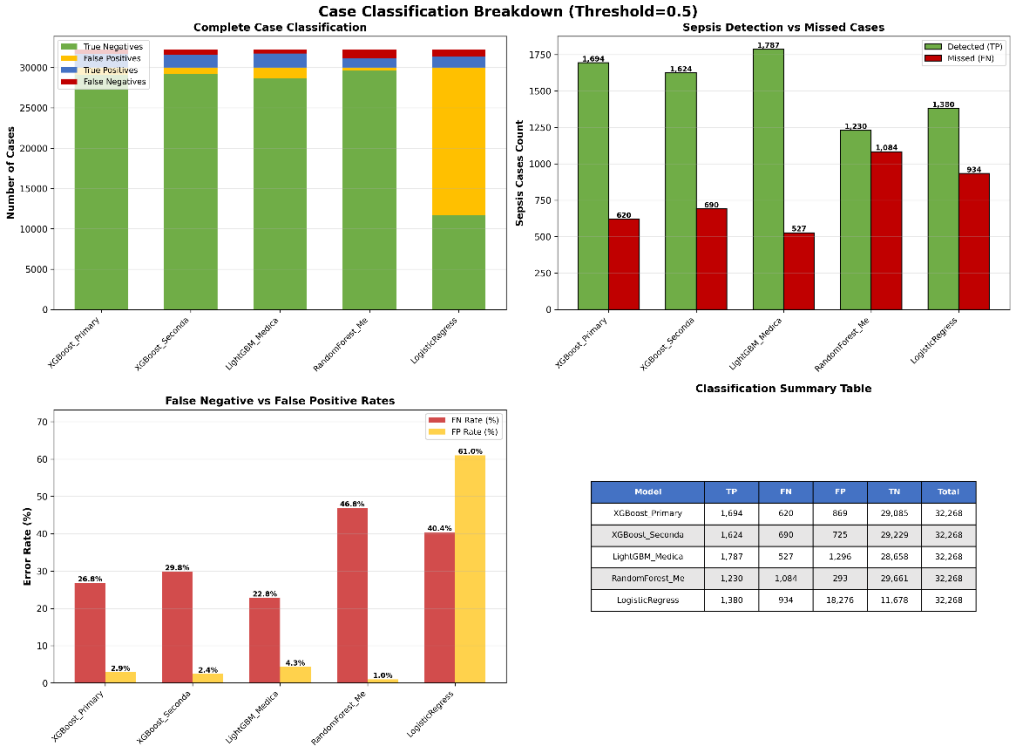


Figura 3: Imagen 5 - Normalized Confusion Matrices (porcentajes por fila).



Model	TP	FN	FP	TN	Total
XGBoost_Primary	1,694	620	869	29,085	32,268
XGBoost_Seconda	1,624	690	725	29,229	32,268
LightGBM_Medica	1,787	527	1,296	28,658	32,268
RandomForest_Me	1,230	1,084	293	29,661	32,268
LogisticRegress	1,380	934	18,276	11,678	32,268

Figura 4: Imagen 8 - Case Classification Breakdown.

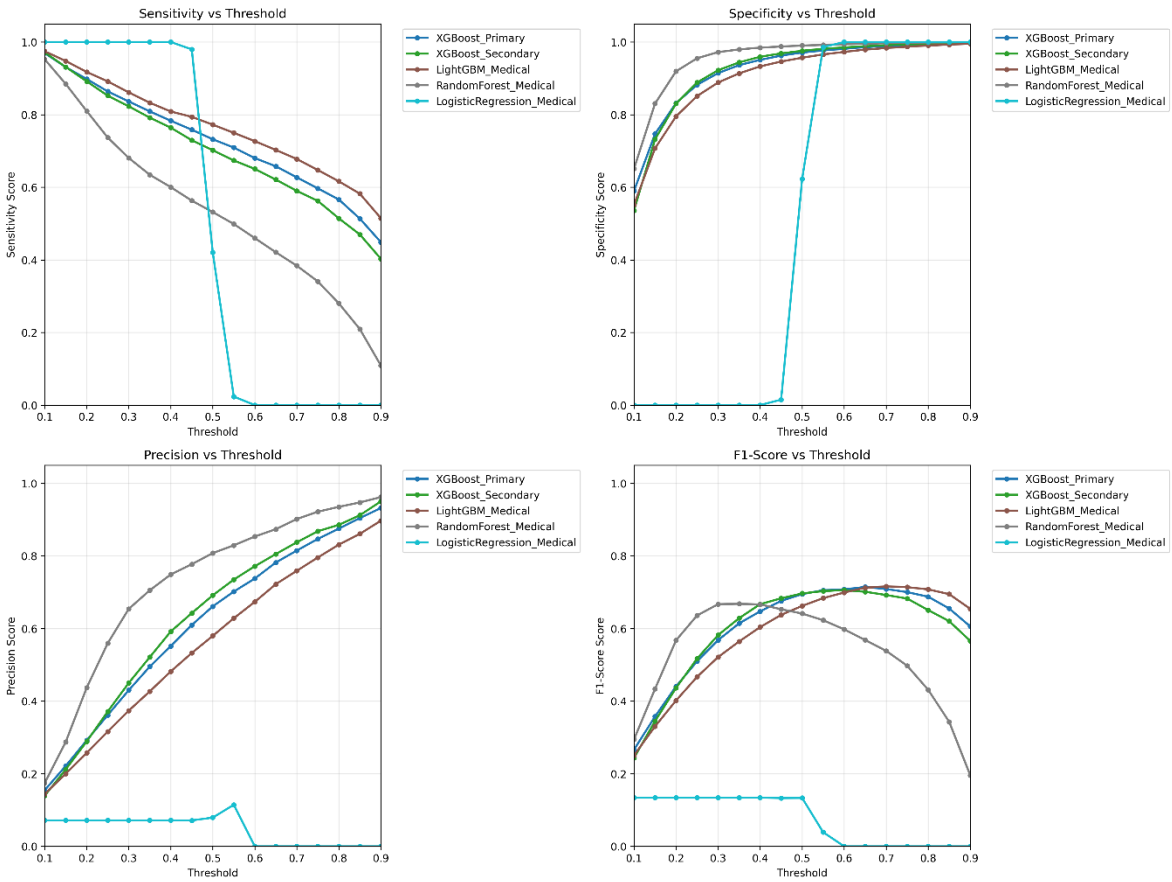


Figura 5: Imagen 6 - Sensitivity/Specificity/Precision/F1 vs Threshold.

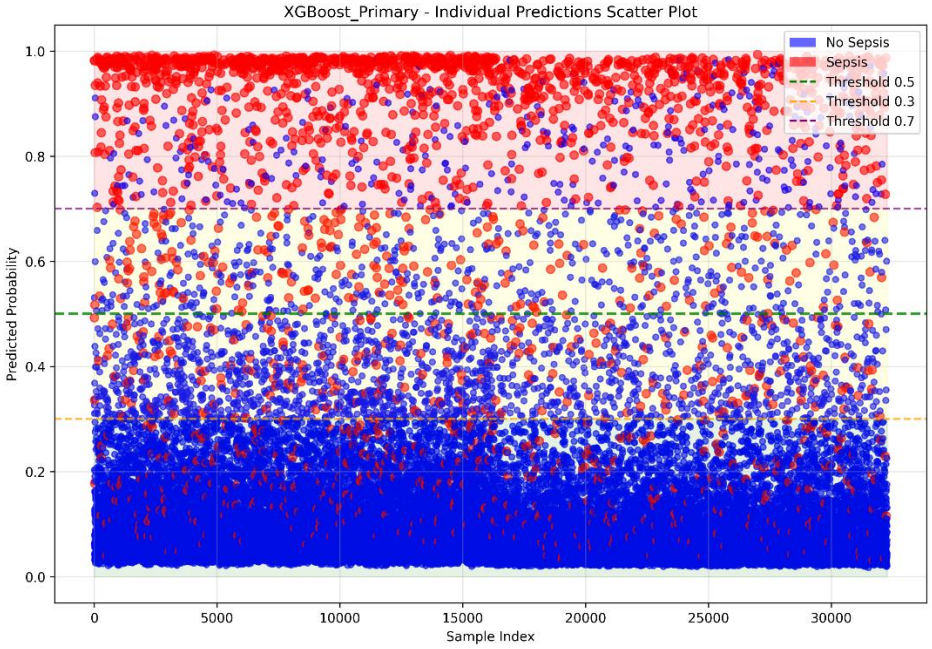


Figura 6: Imagen 4 - XGBoost_Primary Individual Predictions Scatter.

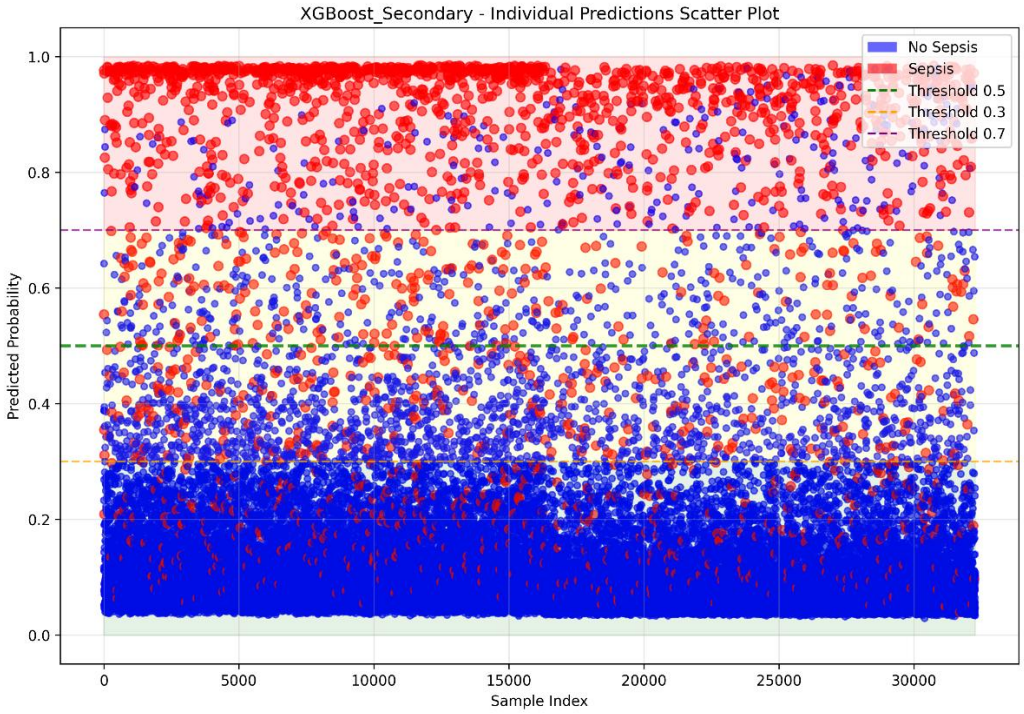


Figura 7: Imagen 3 - XGBoost_Secondary Individual Predictions Scatter

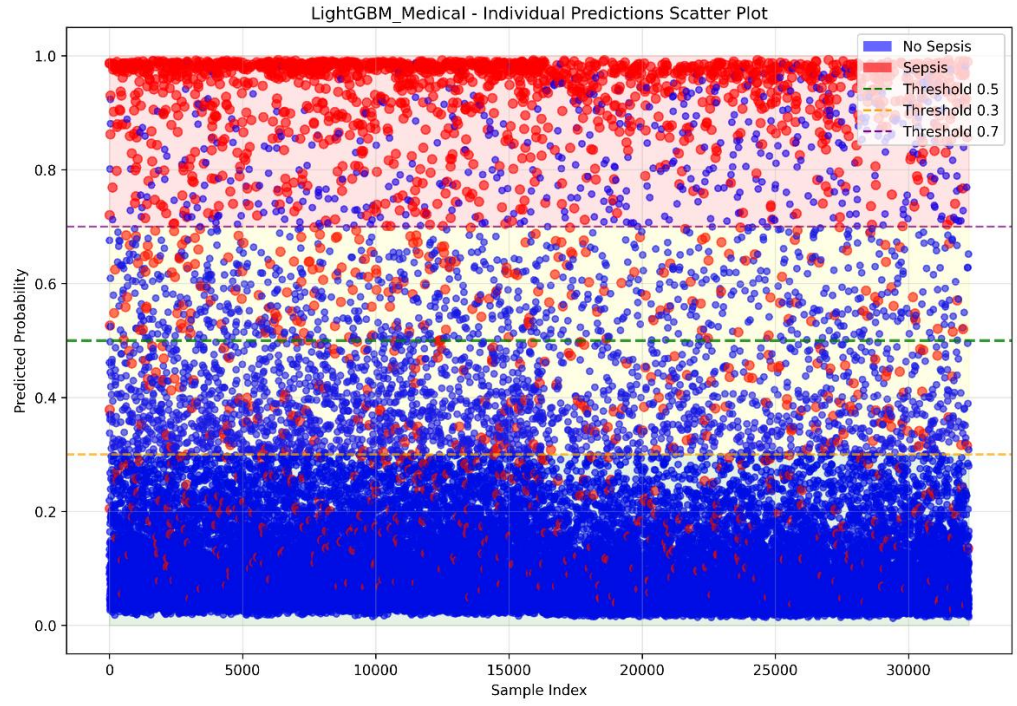


Figura 8: Imagen 2 - LightGBM_Medical Individual Predictions Scatter

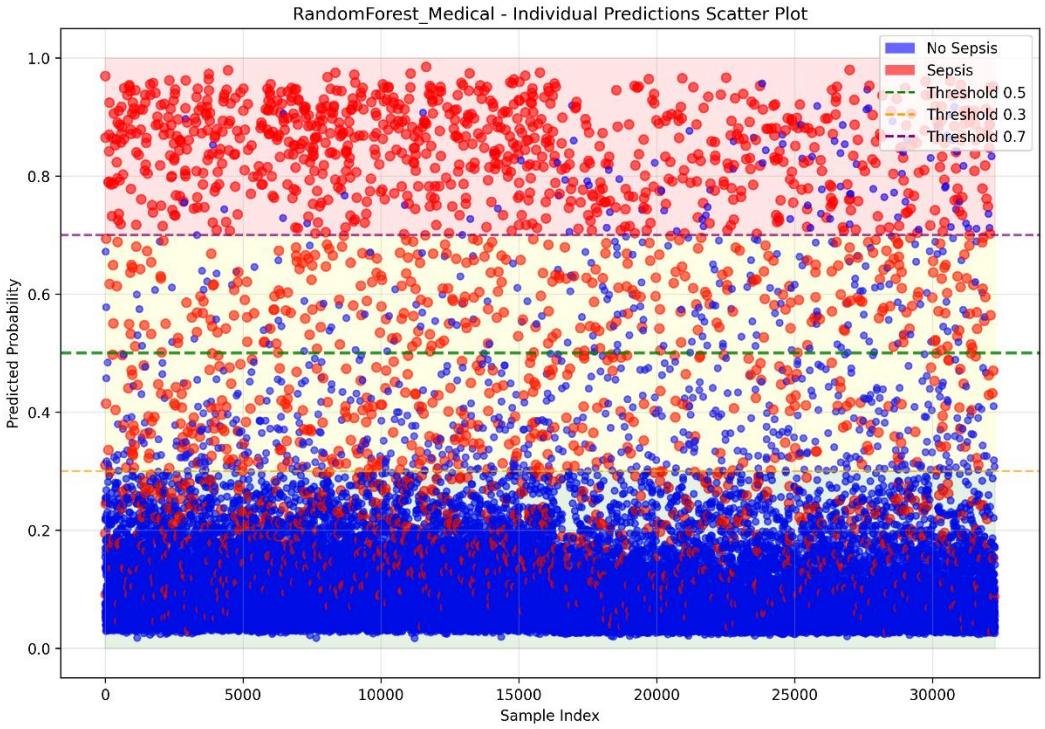


Figura 9: Imagen 1 (segunda) - RandomForest_Medical Individual Predictions Scatter.

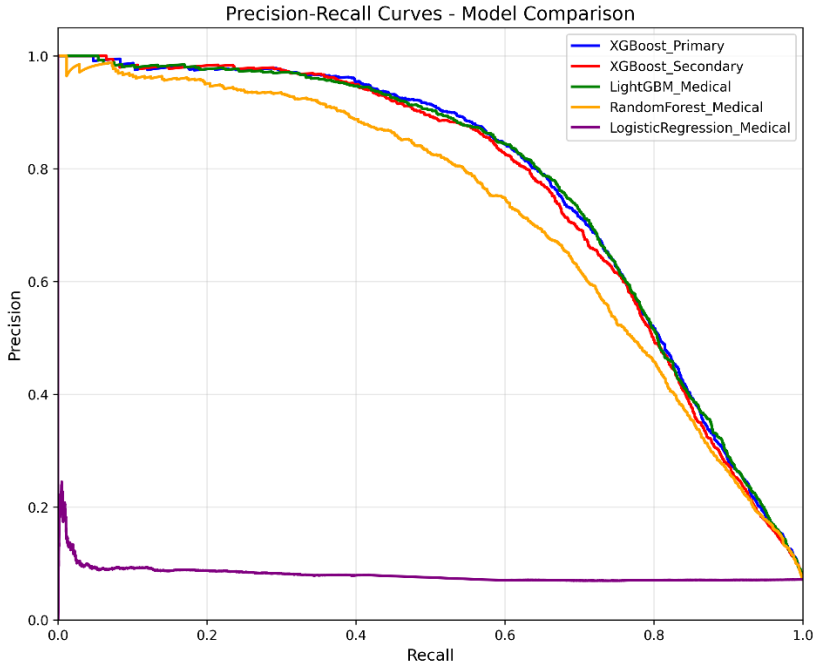


Figura 10: Imagen 9 - Precision-Recall Curves Model Comparison

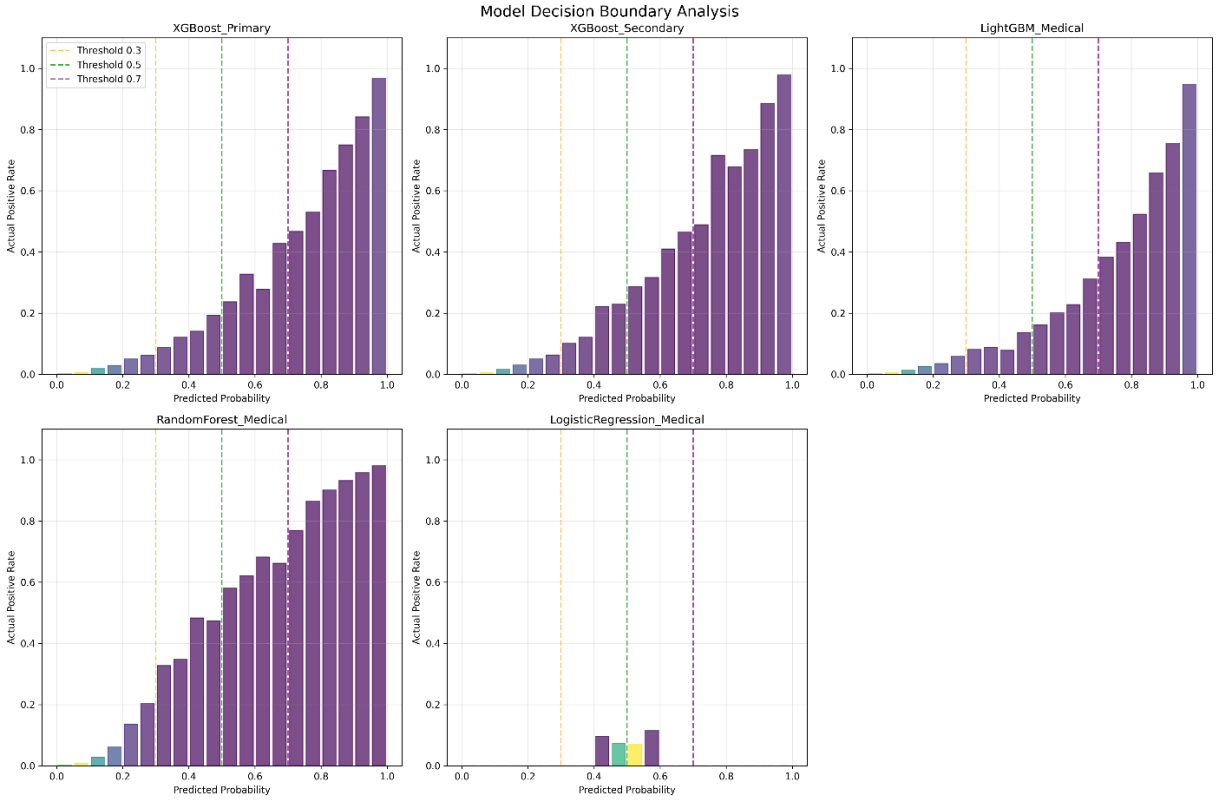


Figura 11: Imagen 7 - Model Decision Boundary Analysis

Bibliografia

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Fleischmann, C., Scherag, A., Adhikari, N. K., Hartog, C. S., Tsaganos, T., Schlattmann, P., Angus, D. C., & Reinhart, K. (2016). Assessment of global incidence and mortality of hospital-treated sepsis: Current estimates and limitations. *American Journal of Respiratory and Critical Care Medicine*, *193*(3), 259-272. https://doi.org/10.1164/rccm.201504-0781OC

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2ª ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la investigación* (6ª ed.). McGraw-Hill Interamericana.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, 3146-3154.

Kumar, A., Roberts, D., Wood, K. E., Light, B., Parrillo, J. E., Sharma, S., Suppes, R., Feinstein, D., Zanotti, S., Taiberg, L., Gurka, D., Kumar, A., & Cheang, M. (2006). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical Care Medicine*, *34*(6), 1589-1596. https://doi.org/10.1097/01.CCM.0000217961.75225.E9

Lever, A., & Mackenzie, I. (2007). Sepsis: Definition, epidemiology, and diagnosis. *BMJ*, *335*(7625), 879-883. https://doi.org/10.1136/bmj.39346.495880.AE

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Rhodes, A., Evans, L. E., Alhazzani, W., Levy, M. M., Antonelli, M., Ferrer, R., Kumar, A., Sevransky, J. E., Sprung, C. L., Nunnally, M. E., Rochwerg, B., Rubenfeld, G. D., Angus, D. C., Annane, D., Beale, R. J., Bellinghan, G. J., Bernard, G. R., Chiche, J. D., Coopersmith, C., ... Dellinger, R. P. (2017). Surviving Sepsis Campaign: International guidelines for management of sepsis and septic

shock: 2016. *Intensive Care Medicine*, *43*(3), 304-377. https://doi.org/10.1007/s00134-017-4683-6

Rudd, K. E., Johnson, S. C., Agesa, K. M., Shackelford, K. A., Tsoi, D., Kievlan, D. R., Colombara, D. V., Ikuta, K. S., Kissoon, N., Finfer, S., Fleischmann-Struzek, C., Machado, F. R., Reinhart, K. K., Rowan, K., Seymour, C. W., Watson, R. S., West, T. E., Marinho, F., Hay, S. I., ... Naghavi, M. (2020). Global, regional, and national sepsis incidence and mortality, 1990–2017: Analysis for the Global Burden of Disease Study. *The Lancet*, *395*(10219), 200-211. https://doi.org/10.1016/S0140-6736(19)32989-7

Seymour, C. W., Liu, V. X., Iwashyna, T. J., Brunkhorst, F. M., Rea, T. D., Scherag, A., Rubenfeld, G., Kahn, J. M., Shankar-Hari, M., Singer, M., Deutschman, C. S., Escobar, G. J., & Angus, D. C. (2016). Assessment of clinical criteria for sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, *315*(8), 762-774. https://doi.org/10.1001/jama.2016.0288

Shimabukuro, D. W., Barton, C. W., Feldman, M. D., Mataraso, S. J., & Das, R. (2017). Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: A randomised clinical trial. *BMJ Open Respiratory Research*, *4*(1), e000234. https://doi.org/10.1136/bmjresp-2017-000234

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J. D., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., van der Poll, T., Vincent, J. L., & Angus, D. C. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, *315*(8), 801-810. https://doi.org/10.1001/jama.2016.0287

World Health Organization. (2020). *Global report on the epidemiology and burden of sepsis: Current evidence, identifying gaps and future directions*. https://www.who.int/publications/i/item/9789240010789

Zhou, Z. H. (2012). *Ensemble methods: Foundations and algorithms*. Chapman and Hall/CRC. https://doi.org/10.1201/b12207