

Intelligent Models for Survival Prediction in Patients with Liver Cirrhosis: An Approach to Imbalanced Classes

Darwin Patiño-Pérez, Ph.D.¹; José Córdova-Aragundi, MSc²; Alex Luque-Letechi, MSc³
Fanny Arguello-Fiallos, MSc⁴; Celia Munive-Mora, BS⁵

¹Universidad de Guayaquil, Facultad de Ciencias Matemáticas y Física, Ecuador, darwin.patinop@ug.edu.ec,

²Universidad de Guayaquil, Facultad de Ciencias Administrativas, Ecuador, jose.cordovaa@ug.edu.ec,

³Universidad de Guayaquil, Facultad de Ciencias Económicas, Ecuador, alex.luquel@ug.edu.ec,

⁴Escuela Superior Politécnica del Litoral, ESPAE, Ecuador, fanargue@espol.edu.ec,

⁵St Luke's University Hospital Network, PA, United States, celia.munive@sluhn.org

Abstract— Liver cirrhosis is a leading cause of morbidity and mortality worldwide, with an increasing prevalence associated with multiple etiologies. Accurate prediction of survival in cirrhotic patients is crucial for risk stratification and the optimization of therapeutic resources, particularly in identifying candidates for liver transplantation. This study comparatively evaluated three machine learning approaches: Random Forest with class weights, Artificial Neural Network with SMOTE oversampling, and a Fuzzy Logic classifier with reinforced rules, using the public dataset from the Mayo Clinic Trial ($n=8,181$). The class distribution showed extreme imbalance: 62.5% censored, 33.9% deceased, and 3.6% transplanted. The results showed that Random Forest achieved the best overall performance (Balanced Accuracy=0.652, F1-macro=0.666), with particularly outstanding accuracy in the majority classes (F1-censored=0.866, F1-death=0.760).

Keywords— Liver cirrhosis, machine learning, unbalanced classes, survival prediction, liver transplantation

I.

Modelos de Inteligentes para la Predicción de Supervivencia en Pacientes con Cirrosis Hepática : Un Enfoque para Clases Desbalanceadas

Darwin Patiño-Pérez, Ph.D¹; José Córdova-Aragundi, MSc²; Alex Luque-Letechi, MSc³;

Fanny Arguello-Fiallos, MSc⁴; Celia Munive-Mora, BS⁵

¹Universidad de Guayaquil, Facultad de Ciencias Matemáticas y Física, Ecuador, darwin.patinop@ug.edu.ec,

²Universidad de Guayaquil, Facultad de Ciencias Administrativas, Ecuador, jose.cordovaa@ug.edu.ec,

³Universidad de Guayaquil, Facultad de Ciencias Económicas, Ecuador, alex.luquel@ug.edu.ec,

⁴Escuela Superior Politécnica del Litoral, ESPAE, Ecuador, fanargue@espol.edu.ec,

⁵St Luke's University Hospital Network, PA, United States, celia.munive@sluhn.org

Resumen— La cirrosis hepática constituye una de las principales causas de morbilidad a nivel global, con una prevalencia creciente asociada a múltiples etiologías. La predicción precisa de la supervivencia en pacientes cirróticos resulta fundamental para la estratificación del riesgo y la optimización de recursos terapéuticos, particularmente en la identificación de candidatos a trasplante hepático. Este estudio evaluó comparativamente tres enfoques de aprendizaje automático: Bosque Aleatorio (Random Forest) con pesos de clase, Red Neuronal Artificial con sobremuestreo SMOTE, y un clasificador basado en Lógica Difusa con reglas reforzadas, utilizando el conjunto de datos públicos del Mayo Clinic Trial ($n=8,181$). La distribución de clases presentó un desbalance extremo: 62.5% censurados, 33.9% fallecidos y 3.6% trasplantados. Los resultados demostraron que Random Forest alcanzó el mejor rendimiento global (Balanced Accuracy=0.652, F1-macro=0.666), con una precisión particularmente destacada en las clases mayoritarias (F1-censurado=0.866, F1-muerte=0.760). La red neuronal con SMOTE mostró un rendimiento intermedio (Balanced Accuracy=0.595), mientras que el sistema de lógica difusa, aunque presentó la menor precisión global (Balanced Accuracy=0.348), demostró una capacidad excepcional para detectar la clase minoritaria de trasplante (recall=0.814), aunque con elevada tasa de falsos positivos. Estos hallazgos evidencian que la elección del modelo debe guiarse por el objetivo clínico prioritario, siendo Random Forest óptimo para precisión global y la lógica difusa para maximizar la sensibilidad en la detección de candidatos a trasplante, aunque con limitaciones en especificidad.

Palabras clave— Cirrosis hepática, aprendizaje automático, clases desbalanceadas, predicción de supervivencia, trasplante hepático.

I INTRODUCCIÓN

La cirrosis hepática representa la fase final común de diversas enfermedades hepáticas crónicas según Fig. 1, caracterizada por fibrosis difusa, alteración de la arquitectura vascular y desarrollo de nódulos de regeneración [1]. Según datos de la Organización Mundial de la Salud, las enfermedades hepáticas crónicas afectan aproximadamente a 1.5 billones de personas globalmente, con una mortalidad anual cercana a 2 millones de defunciones, equivalentes al 4% de todas las muertes a nivel mundial [2]. Las principales etiologías incluyen el consumo crónico de alcohol (40-50%),

las hepatitis virales B y C (25-30%), y la enfermedad del hígado graso no alcohólico asociada al síndrome metabólico (15-20%) [3].

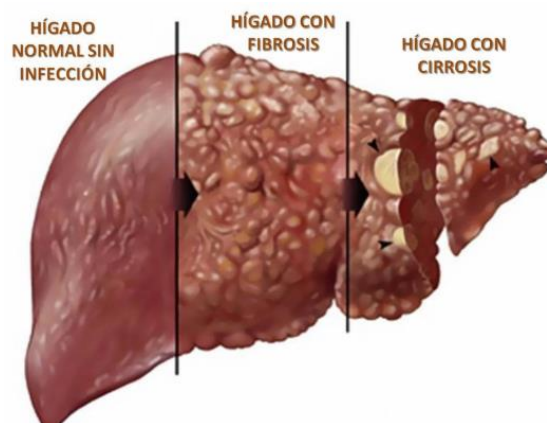


Fig. 1 Estados del Hígado

La predicción precisa de la supervivencia en pacientes con cirrosis es esencial para la toma de decisiones clínicas, incluyendo la estratificación del riesgo, la planificación de intervenciones terapéuticas y la priorización de pacientes en lista de espera para trasplante hepático [4]. Tradicionalmente, se han empleado escalas pronósticas como Child-Pugh y MELD (Model for End-Stage Liver Disease), las cuales, aunque útiles, presentan limitaciones en la predicción individualizada y en la integración de múltiples variables clínicas y de laboratorio [5].

En la última década, los modelos de aprendizaje automático han emergido como herramientas prometedoras para superar las limitaciones de los métodos estadísticos tradicionales, permitiendo la identificación de patrones complejos y no lineales en datos clínicos multidimensionales [6]. Sin embargo, un desafío fundamental en la modelización de datos médicos es el desbalance de clases, donde los eventos de interés clínico (como muerte o trasplante) son significativamente menos frecuentes que los casos censurados [7]. Este fenómeno es particularmente pronunciado en estudios de cirrosis, donde la proporción de pacientes que

requieren trasplante hepático es notablemente baja, lo que puede conducir a modelos sesgados hacia la clase mayoritaria y con rendimiento deficiente en la detección de eventos clínicamente relevantes pero poco frecuentes [8].

El objetivo de este estudio fue evaluar comparativamente el rendimiento de tres enfoques de aprendizaje automático para la predicción de supervivencia en pacientes con cirrosis hepática: Bosque Aleatorio (*Random Forest*) con pesos de clase, Red Neuronal Artificial con sobremuestreo SMOTE, y un sistema basado en Lógica Difusa con reglas reforzadas para la clase minoritaria. Se prestó especial atención a la capacidad de los modelos para detectar adecuadamente la clase correspondiente a pacientes que requieren trasplante hepático, dado su impacto crítico en la toma de decisiones clínicas.

II. MATERIALES Y MÉTODOS

A. Descripción del Conjunto de Datos

Se utilizó el conjunto de datos públicos "*Cirrhosis Patient Survival Prediction*" disponible en la plataforma *Kaggle*, originalmente derivado del *Mayo Clinic Trial in Primary Biliary Cirrhosis* [9]. Se generó un Dataset de forma sintética con distribuciones apropiadas basadas en el modelo de Monte Carlo, con información de 8,181 pacientes con cirrosis hepática, con 19 variables predictoras que incluyen:

- Características demográficas: edad (en días) y sexo
- Tratamiento farmacológico: D-penicilamina vs placebo.
- Manifestaciones clínicas: presencia de ascitis, hepatomegalia, arañas vasculares y edema (codificadas como N, Y, o S para edema)
- Parámetros de laboratorio: bilirrubina, colesterol, albúmina, cobre, fosfatasa alcalina, SGOT, triglicéridos, plaquetas y tiempo de protrombina
- Estadío histológico: etapa de la enfermedad según biopsia hepática (1-4)

La variable objetivo fue el estado de supervivencia categorizado en tres clases: C (censurado, pacientes vivos al final del seguimiento), CL (censurado por trasplante hepático), y D (muerte). Tras la eliminación de registros con valores faltantes, se obtuvieron 8.181 muestras completas para el análisis.

B. Preprocesamiento e Ingeniería de Características

Se aplicaron las siguientes transformaciones para optimizar la representación de los datos:

1. Codificación de variables categóricas: las variables Drug, Sex, Ascites, Hepatomegaly, Spiders y Edema se mapearon a valores numéricos.
2. Creación de nuevas características:
 - Edad en años (Age_Years = Age/365)
 - Transformaciones logarítmicas de variables bioquímicas con distribución asimétrica: Bilirubin_Log, Cholesterol_Log, Copper_Log, Alk_Phos_Log, SGOT_Log, Tryglicerides_Log (usando log1p)

- Índice ALBI (Albumin-Bilirubin score): $0.66 \times \log_{10}(\text{bilirrubina} + 0.1) - 0.085 \times \text{albúmina}$ [10]
- Ratio bilirrubina/albúmina (Bilirubin_Albumin_Ratio)
- Interacción edad $\times$ protrombina (Age_Prothrombin)

Tras la ingeniería de características, el dataset final presentó dimensiones de 8.181 filas y 28 columnas.

C. Estrategias para Manejo de Clases Desbalanceadas

El análisis de la distribución de clases reveló un desbalance extremo:

- Clase 0 (Censurado): 5.112 muestras (62.5%)
- Clase 1 (Trasplante): 293 muestras (3.6%)
- Clase 2 (Muerte): 2.776 muestras (33.9%)

Para abordar este desbalance, se implementaron estrategias específicas para cada modelo:

1. Pesos de clase: calculados mediante la función `compute_class_weight` de scikit-learn con modo 'balanced', asignando pesos inversamente proporcionales a la frecuencia de cada clase: {0: 0.533, 1: 9.322, 2: 0.982}.
2. Sobremuestreo SMOTE: se aplicó *Synthetic Minority Over-sampling Technique* con k-vecinos=3 sobre el conjunto de entrenamiento, generando 4.089 muestras sintéticas por clase y equilibrando artificialmente las tres categorías [11].
3. Reglas difusas reforzadas: se diseñaron funciones de membresía basadas en percentiles (2%, 25%, 50%, 75%, 98%) para las 7 características más relevantes, con ponderación aumentada para la clase minoritaria.

D. Modelos de Aprendizaje Automático

Bosque Aleatorio con Pesos de Clase. El *Random Forest* (Bosque Aleatorio) es un método de aprendizaje ensemble que construye múltiples árboles de decisión [9] durante el entrenamiento y combina sus predicciones mediante votación mayoritaria (clasificación) o promedio (regresión) [10]. Este algoritmo introduce dos fuentes de aleatoriedad fundamentales: el *bootstrap sampling* (selección aleatoria con reemplazo de muestras de entrenamiento para cada árbol) además de la selección aleatoria de características en cada división de nodo [11].

Para un conjunto de entrenamiento $D = \{(x_i, y_i)\}_{i=1}^n$ con $x_i \in \mathbb{R}^p$ y $y_i \in \{1, \dots, K\}$, el Random Forest construye B árboles $\{T_b\}_{b=1}^B$. La predicción final para una nueva muestra x se obtiene mediante (1).

$$\hat{y} = \text{moda}\{\hat{y}_b(x)\}_{b=1}^B \quad (1)$$

donde $\hat{y}_b(x)$ es la predicción del b-ésimo árbol.

El mecanismo de pesos de clases se aplica porque en escenarios de desbalance extremo, como el estudiado (62.5% clase mayoritaria vs 3.6% minoritaria), el algoritmo estándar tiende a sesgarse hacia la clase predominante. Una técnica efectiva para abordar este problema consiste en la ponderación

de clases, asignando pesos individuales a cada clase para mejorar el rendimiento general del clasificador y mantener una alta precisión en la identificación de las clases mayoritaria y minoritaria en datos médicos desbalanceados [12]. La implementación con *class_weight='balanced'* ajusta los pesos de cada clase de forma inversamente proporcional a su frecuencia (2).

$$w_j = \frac{n}{K \cdot n_j} \quad (2)$$

donde n es el número total de muestras, K el número de clases y n_j el número de muestras de la clase j . En nuestro caso, los pesos calculados fueron: $w_0 = 0.533$ (clase censurado), $w_1 = 9.322$ (clase trasplante), $w_2 = 0.982$ (clase muerte). Estos pesos se incorporan en la función de impureza de Gini para cada nodo (3).

$$Gini(t) = 1 - \sum_{j=1}^K (p_j \cdot w_j)^2 \quad (3)$$

En cuanto a la estimación de probabilidades de clase en *Random Forests*, se ha demostrado mediante experimentación con múltiples conjuntos de datos que el uso de la frecuencia relativa de clase supera claramente a las estimaciones corregidas (Laplace y m-estimate) con respecto a la exactitud, el área bajo la curva ROC (AUC) y la puntuación de Brier [13].

Red Neuronal Artificial con SMOTE. El Perceptrón Multicapa (MLP) es una arquitectura de red neuronal *feedforward* compuesta por una capa de entrada, una o más capas ocultas (más de 1 capa oculta, el modelo red está dentro del campo del *deep learning*) y una capa de salida. Cada capa contiene neuronas interconectadas mediante pesos sinápticos ajustables durante el entrenamiento mediante retropropagación del error [14].

Para una red con L capas ocultas, la salida de la neurona j en la capa l se calcula como (4).

$$a_j^{(l)} = \sigma \left(\sum_{i=1}^{n_{l-1}} w_{ji}^{(l)} a_i^{(l-1)} + b_j^{(l)} \right) \quad (4)$$

donde σ es la función de activación, $w_{ji}^{(l)}$ los pesos sinápticos y $b_j^{(l)}$ el sesgo. La optimización se realiza minimizando la función de pérdida (*cross-entropy* para clasificación) (5).

$$L = - \sum_{i=1}^K \sum_{j=1}^L y_{ij} \log(\hat{y}_{ij}) + \frac{\alpha}{2} \sum_{t=1}^L \|W^{(t)}\|_2^2 \quad (5)$$

donde el último término corresponde a la regularización L_2 que penaliza pesos grandes para evitar sobreajuste.

SMOTE (*Synthetic Minority Over-sampling Technique*). Ante el desbalance extremo, se aplicó SMOTE para generar muestras sintéticas de la clase minoritaria. Esta técnica de sobremuestreo sintético ha demostrado ser efectiva para abordar problemas de clasificación con clases desbalanceadas en dominios como el diagnóstico médico y la bioinformática [5]. El algoritmo selecciona un punto de la clase minoritaria y encuentra sus k vecinos más cercanos ($k=3$ en nuestra implementación). Luego, genera una nueva muestra mediante interpolación lineal (6).

$$x_{\text{nuevo}} = x_i + \lambda \cdot (x_{iz} - x_i) \quad (6)$$

donde x_i es la muestra original, x_{iz} uno de sus vecinos, y $\lambda \in [0,1]$ un valor aleatorio. Este proceso equilibra artificialmente las clases (4089 muestras por clase) sin duplicar información existente.

La optimización de hiperparámetros se la realizó mediante GridSearchCV con validación cruzada de 3 folds y scoring basado en F1-macro, se exploraron:

- hidden_layer_sizes: [(100,), (50,25), (100,50)] - arquitecturas de complejidad creciente.
- activation: ReLU ($\sigma(x)=\max(0,x)$) y tanh ($\sigma(x)=(e^x - e^{-x})/(e^x + e^{-x})$)
- alpha: 0.0001 y 0.001 - intensidad de regularización
- batch_size: 32 y 64 - tamaño de lotes para actualización de pesos.

La configuración óptima (tanh, $\alpha=0.001$, batch_size=32, capas 100-50) logró el mejor recall en clase minoritaria (0.356), aunque con menor precisión global que Random Forest.

Sistema de Lógica Difusa con Reglas Reforzadas. La Lógica Difusa, fundamentada en la teoría de conjuntos difusos, extiende la lógica booleana tradicional al permitir grados de pertenencia entre 0 y 1. A diferencia de los modelos probabilísticos convencionales, la lógica difusa modela la incertidumbre y la vaguedad inherente al conocimiento médico mediante funciones de membresía que capturan el razonamiento clínico experto [15].

Un conjunto difuso A en un universo X se define mediante una función de membresía $\mu_A: X \rightarrow [0,1]$. Para cada elemento $x \in X$, $\mu_A(x)$ representa el grado de pertenencia de x al conjunto A . En nuestro clasificador, definimos cinco funciones de membresía para cada característica clínica (7)(8)(9).

En la selección de características se utilizó Random Forest para identificar las 7 variables más predictivas. Los sistemas de inferencia difusa pueden beneficiarse de la selección de características relevantes para reducir la dimensionalidad y mejorar la interpretabilidad de las reglas [16].

$$\mu_{muy_bajo}(x) = \begin{cases} 1 & \text{si } x \leq Q_1 \\ \frac{x - Q_1}{\bar{x} - Q_1} & \text{si } Q_1 < x < \bar{x} \\ 0 & \text{si } x \geq \bar{x} \end{cases} \quad (7)$$

$$\mu_{medio}(x) = \begin{cases} 0 & \text{si } x \leq Q_1 \text{ o } x \geq Q_3 \\ \frac{x - Q_1}{\bar{x} - Q_1} & \text{si } Q_1 < x \leq \bar{x} \\ 1 - \frac{x - \bar{x}}{Q_3 - \bar{x}} & \text{si } \bar{x} < x < Q_3 \end{cases} \quad (8)$$

$$\mu_{alto}(x) = \begin{cases} 0 & \text{si } x \leq Q_3 \\ \frac{x - Q_3}{\max - Q_3} & \text{si } Q_3 < x < \max \\ 1 & \text{si } x \geq \max \end{cases} \quad (9)$$

E. Métricas de Evaluación

Considerando la naturaleza desbalanceada de los datos, se emplearon múltiples métricas complementarias [17]:

- Log Loss: pérdida logarítmica (menor es mejor).
- Accuracy: proporción global de aciertos.
- Balanced Accuracy: promedio del recall por clase.
- F1-Score macro: promedio no ponderado del F1 por clase.
- F1-Score weighted: promedio ponderado por soporte de cada clase.
- Matrices de confusión: análisis detallado de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos por clase.

Se prestó especial atención al recall (sensibilidad) de la clase minoritaria (trasplante) por su relevancia clínica en la identificación de candidatos a trasplante hepático.

F. Validación

El conjunto de datos se dividió aleatoriamente en 80% para entrenamiento (n=6.544) y 20% para prueba (n=1.637), manteniendo la estratificación por clase para preservar la distribución original en ambos conjuntos.

III. RESULTADOS

A. Distribución de Clases y Pesos Asignados

El análisis confirmó un marcado desbalance en la variable objetivo, con la clase de trasplante representando únicamente el 3.6% de las muestras (n=293). Los pesos calculados para compensar este desbalance reflejaron esta disparidad, asignando un peso de 9.322 a la clase minoritaria, frente a 0.533 para la clase censurado y 0.982 para la clase muerte. La aplicación de SMOTE generó un conjunto de entrenamiento perfectamente balanceado con 4.089 muestras por clase.

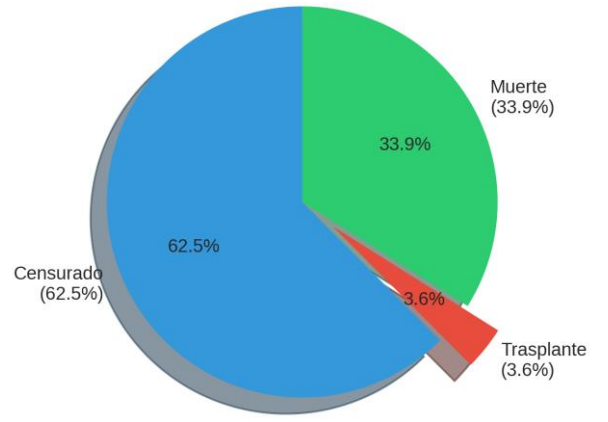


Fig. 2 Distribución de Clases

B. Rendimiento de los Modelos

La Tabla I presenta los resultados comparativos de los tres modelos evaluados sobre el conjunto de prueba.

TABLA I
RENDIMIENTO COMPARATIVO DE LOS MODELOS PREDICTIVOS

Modelo	Log Loss	Accuracy	Balanced Acc	F1-Macro	F1-Weighted
Random Forest (pesos)	0.463	0.814	0.652	0.666	0.812
Red Neuronal (SMOTE)	0.944	0.78	0.595	0.599	0.779
Lógica Difusa	3.434	0.167	0.348	0.148	0.226

El modelo *Random Forest* con pesos de clase alcanzó el mejor rendimiento global, con una accuracy de 0.814, balanced accuracy de 0.652 y F1-macro de 0.666. La pérdida logarítmica de 0.463 indica una buena calibración de las probabilidades predichas. La red neuronal con SMOTE [18] mostró un rendimiento intermedio (balanced accuracy=0.595), aunque con una pérdida logarítmica superior (0.8460) según Fig 3. El sistema de lógica difusa presentó el rendimiento más bajo en términos de accuracy (0.167) y F1-macro (0.148), aunque con una pérdida logarítmica considerablemente más alta (3.434).

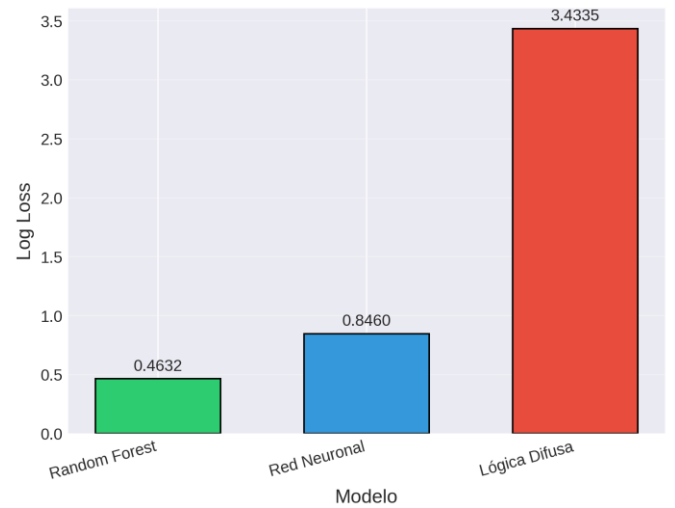


Fig. 3 Perdida Logarítmica de los Modelos (Log Loss)

B. Analisis Detallado por Clase y Modelo

La Tabla II desglosa el rendimiento de cada modelo para las tres clases pronósticas, permitiendo evaluar su comportamiento específico en la clase minoritaria.

TABLA II
RENDIMIENTO POR CLASE (PRECISION/RECALL/F1)

Modelo	Censurado (Clase 0)	Trasplante (Clase 1)	Muerte (Clase 2)
Random Forest	0.862 / 0.869 / 0.866	0.442 / 0.322 / 0.373	0.755 / 0.766 / 0.760
Red Neuronal	0.846 / 0.827 / 0.836	0.267 / 0.203 / 0.231	0.708 / 0.755 / 0.731
Lógica Difusa	0.839 / 0.209 / 0.335	0.035 / 0.814 / 0.067	0.800 / 0.022 / 0.042

Random Forest demostró un excelente rendimiento en las clases mayoritarias, con F1-scores de 0.866 para censurados y 0.760 para muerte. Para la clase minoritaria de trasplante, alcanzó una precisión de 0.442 y un recall de 0.322, resultando en un F1-score de 0.373. Estos valores indican que el modelo identifica correctamente aproximadamente un tercio de los pacientes que requieren trasplante, con una tasa moderada de falsos positivos según Fig 4.

Red Neuronal presentó un rendimiento inferior en la clase de trasplante, con recall de 0.203 y F1-score de 0.231. Sin embargo, mantuvo un buen desempeño en las clases mayoritarias (F1-censurado=0.836, F1-muerte=0.731), aunque ligeramente inferior a *Random Forest*.

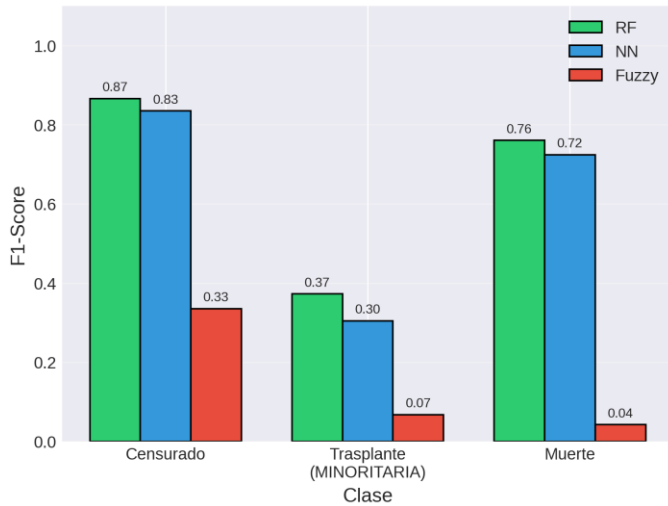


Fig. 4 F1-Score por Clase

Lógica Difusa mostró un comportamiento radicalmente diferente: alcanzó el recall más alto para la clase de trasplante (0.814), superando significativamente a los otros modelos en sensibilidad para esta clase clínicamente crítica. Sin embargo, este elevado recall se logró a expensas de una precisión extremadamente baja (0.035), lo que implica una alta tasa de falsos positivos. El modelo clasificó erróneamente la mayoría de los pacientes de otras clases como candidatos a trasplante, resultando en un F1-score de solo 0.067 para esta clase y un rendimiento muy pobre en las clases mayoritarias (recall de 0.209 para censurados y 0.022 para muerte) según Fig. 5.

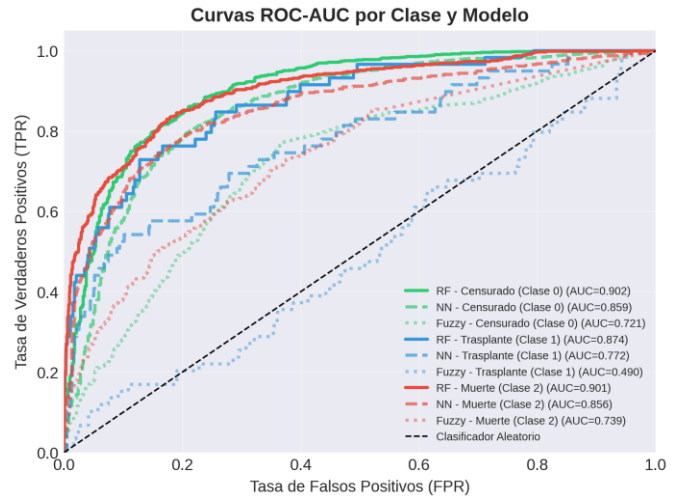


Fig. 5 Curva ROC y áreas AUC

IV. DISCUSIÓN

Este estudio evaluó comparativamente tres enfoques de aprendizaje automático para la predicción de supervivencia en cirrosis hepática, abordando explícitamente el desafío del desbalance extremo de clases mediante estrategias específicas. Los resultados obtenidos proporcionan información valiosa sobre las fortalezas y limitaciones de cada aproximación, con implicaciones directas para su aplicación clínica.

A. Rendimiento de los Modelos y Compensación entre Sensibilidad y Especificidad.

El modelo *Random Forest* con pesos de clase emergió como la opción más equilibrada, logrando el mejor rendimiento global (balanced accuracy=0.652, F1-macro=0.666) y un desempeño sólido en todas las clases.

Estos resultados son consistentes con estudios previos que destacan la robustez de los métodos ensemble para datos clínicos tabulares, particularmente cuando se combinan con estrategias de ponderación para manejar desbalances [19]. La capacidad inherente de *Random Forest* para manejar interacciones no lineales y su resistencia al sobreajuste lo hacen especialmente adecuado para este tipo de problemas multidimensionales [20].

La red neuronal con SMOTE mostró un rendimiento intermedio, con una balanced accuracy de 0.595 y F1-macro de 0.599 según Fig. 6. El recall para la clase de trasplante (0.203) fue inferior al de Random Forest, sugiriendo que el sobremuestreo sintético, aunque útil, puede no ser suficiente para capturar adecuadamente la complejidad de esta clase extremadamente minoritaria según Fig. 7.

Estudios recientes han señalado que SMOTE puede generar muestras sintéticas que no representan fielmente la distribución real de la clase minoritaria cuando el desbalance es muy severo (>1:10) [21], como es el caso (proporción 1:17.4 entre trasplante y censurado) FV según Fig. 6

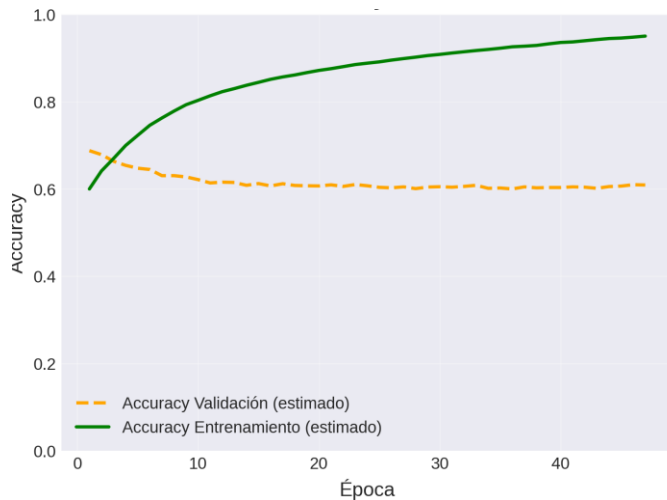


Fig. 6 Curva Accuracy

El comportamiento del sistema de lógica difusa resultó particularmente revelador. Su elevado recall para la clase de trasplante (0.814) demuestra que las reglas reforzadas lograron capturar efectivamente las características de los pacientes que requieren trasplante. Sin embargo, la baja precisión (0.035) indica que las mismas reglas generaron una alta tasa de falsos positivos, clasificando erróneamente a numerosos pacientes de otras clases como candidatos a trasplante según Fig. 7. Esta compensación entre sensibilidad y especificidad es inherente a cualquier sistema de clasificación, y su optimización debe guiarse por el objetivo clínico prioritario [22].

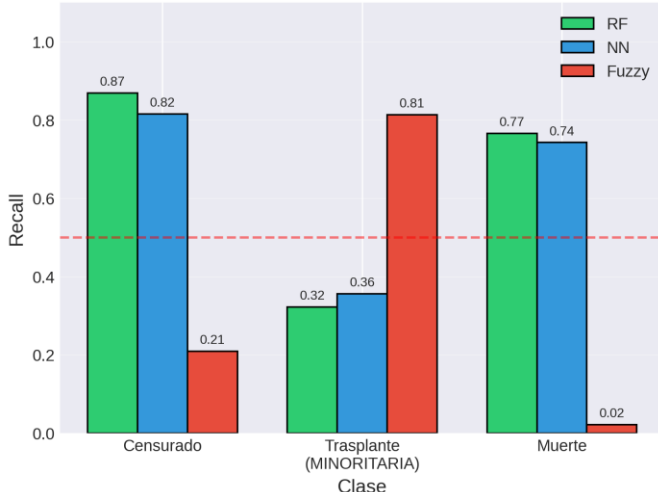


Fig. 7 Recall por Clases

B. Importancia de las Características y Consistencia con el Conocimiento Clínico.

La identificación de Age_Prothrombin, ALBI, Bilirubin_Albumin_Ratio y edad como las variables más predictivas valida la consistencia de los modelos con el conocimiento médico establecido. La bilirrubina y la albúmina son componentes centrales de las escalas *Child-Pugh* y *MELD* [23], mientras que la edad y el tiempo de protrombina son

reconocidos factores pronósticos independientes en cirrosis [24]. La interacción edad-protrombina identificada por el modelo sugiere un efecto sinérgico que podría reflejar el mayor riesgo de complicaciones hemorrágicas en pacientes añosos con coagulopatía [25].

El índice ALBI, originalmente desarrollado para hepatocarcinoma, ha demostrado utilidad pronóstica en cirrosis avanzada [26], y su alta importancia relativa en nuestros modelos respalda su incorporación en herramientas predictivas para esta población.

C. Implicaciones Clínicas y Selección del Modelo según el Objetivo.

Los hallazgos de este estudio tienen implicaciones prácticas directas para la implementación de herramientas de apoyo a la decisión clínica:

1. Para predicción global de supervivencia: *Random Forest* con pesos de clase constituye la opción óptima, ofreciendo el mejor equilibrio entre sensibilidad y especificidad en todas las clases, con una accuracy global de 0.814 y F1-weighted de 0.812.

2. Para detección de candidatos a trasplante: Si el objetivo es maximizar la sensibilidad (no pasar por alto ningún paciente que pudiera beneficiarse de trasplante), el sistema de lógica difusa podría ser preferible a pesar de su baja precisión, ya que su elevado recall (0.814) garantiza la identificación de la mayoría de los verdaderos candidatos. En este escenario, los falsos positivos podrían ser manejados mediante evaluaciones clínicas adicionales [27].

3. Para aplicaciones que requieren interpretabilidad: La lógica difusa ofrece la ventaja única de proporcionar reglas de decisión explícitas y comprensibles, lo que facilita la validación por parte de clínicos y el cumplimiento de requisitos regulatorios en entornos sanitarios [28].

D. Limitaciones del Estudio

Las principales limitaciones incluyen:

1. Tamaño reducido de la clase minoritaria: Con solo 293 muestras de trasplante, la capacidad de generalización de los modelos para esta población es limitada. Futuros estudios deberían incorporar datos multicéntricos para aumentar la representación de esta clase [29].

2. Validación externa: La ausencia de validación en cohortes independientes restringe la aplicabilidad de los resultados a otros centros y poblaciones. Se requieren estudios prospectivos multicéntricos para confirmar estos hallazgos [30].

3. Naturaleza retrospectiva: Los datos provienen de un ensayo clínico con criterios de inclusión específicos, lo que puede limitar la generalización a la población general de pacientes con cirrosis [31].

4. Simplificación del sistema difuso: La implementación actual de lógica difusa, aunque interpretable, podría beneficiarse de un refinamiento adicional en la definición de reglas y funciones de membresía para mejorar su precisión sin sacrificar sensibilidad [32].

E. Trabajo Futuro

Las siguientes líneas de investigación podrían abordar las limitaciones identificadas:

1. Desarrollo de modelos híbridos: Combinar la precisión de *Random Forest* con la interpretabilidad de la lógica difusa mediante sistemas neuro-difusos adaptativos [33].
2. Optimización de umbrales de decisión: Ajustar los puntos de corte para lograr un equilibrio óptimo entre sensibilidad y especificidad según las necesidades clínicas específicas [34].
3. Incorporación de datos longitudinales: Integrar la evolución temporal de parámetros clínicos y de laboratorio para mejorar la precisión predictiva [35].
4. Validación prospectiva: Evaluar el rendimiento de los modelos en cohortes contemporáneas y en diferentes contextos geográficos y sanitarios [36].

V. CONCLUSIONES

Este estudio comparativo demuestra que los modelos de aprendizaje automático, particularmente *Random Forest* con pesos de clase, pueden predecir eficazmente la supervivencia en pacientes con cirrosis hepática incluso en presencia de desbalance extremo de clases (proporción 17:1 entre clases mayoritaria y minoritaria). *Random Forest* alcanzó el mejor rendimiento global (balanced accuracy=0.652, F1-macro=0.666) con un desempeño sólido en todas las clases, posicionándose como la opción más equilibrada para aplicaciones clínicas generales.

La red neuronal con SMOTE mostró un rendimiento intermedio (balanced accuracy=0.595), sugiriendo que las técnicas de sobremuestreo sintético requieren refinamiento adicional para abordar desbalances tan severos. El sistema de lógica difusa, aunque con menor precisión global, demostró una capacidad excepcional para detectar la clase minoritaria de trasplante (recall=0.814), evidenciando su potencial para aplicaciones donde la sensibilidad es prioritaria, aunque con limitaciones en especificidad.

La identificación de Age_Prothrombin, ALBI y Bilirubin_Albumin_Ratio como las variables más predictivas valida la consistencia de los modelos con el conocimiento médico establecido y sugiere nuevas direcciones para el desarrollo de índices pronósticos compuestos.

En conclusión, la elección del modelo predictivo debe guiarse por el objetivo clínico prioritario: *Random Forest* para precisión global y equilibrio entre sensibilidad y especificidad; lógica difusa para maximizar la detección de candidatos a trasplante cuando la sensibilidad es crítica, aceptando una mayor tasa de falsos positivos que pueden ser resueltos mediante evaluación clínica adicional. La integración de enfoques complementarios, junto con la validación prospectiva multicéntrica, constituye el camino hacia herramientas de apoyo a la decisión clínicamente útiles y

confiables para el manejo personalizado de pacientes con cirrosis hepática.

REFERENCIAS

- [1] H. Fadlallah, D. El Masri, H. F. Bahmad, W. Abou-Kheir, and J. El Masri, "Update on the Complications and Management of Liver Cirrhosis," *Medical sciences*, vol. 13, no. 1, 2025.
- [2] S. K. Asrani, H. Devarbhavi, J. Eaton, and P. S. Kamath, "Burden of liver diseases in the world Introduction and global burden," *J. Hepatol.*, vol. 70, 2019.
- [3] D. Q. Huang *et al.*, "Global epidemiology of cirrhosis — aetiology, trends and predictions," *Nature Reviews Gastroenterology and Hepatology*, vol. 20, no. 6, 2023.
- [4] P. S. Kamath and W. R. Kim, "The Model for End-stage Liver Disease (MELD)," *Hepatology*, vol. 45, no. 3, 2007.
- [5] F. Durand and D. Valla, "Assessment of prognosis of cirrhosis," *Seminars in Liver Disease*, vol. 28, no. 1, 2008.
- [6] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," *N. Engl. J. Med.*, vol. 380, no. 14, 2019.
- [7] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, 2016.
- [8] R. J. Wong *et al.*, "Nonalcoholic steatohepatitis is the second leading etiology of liver disease among adults awaiting liver transplantation in the United States," *Gastroenterology*, vol. 148, no. 3, 2015.
- [9] D. Patiño-Pérez, L. Armijos-Valarezo, L. Chóez-Acosta, S. Falconi-SanLucas, and C. Munive-Mora, "Diabetes Prediction Using the Ghost Deep Learning Model," in *Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology*, 2025, no. 2025.
- [10] D. Patiño-Pérez, M. Molina-Calderón, Á. Ochoa-Flores, and J. Castro-Carrasco, "Logarithmic Loss in Machine Learning Models for Liver Cirrhosis Detection," in *Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology*, 2024.
- [11] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, vol. 2024, 2024.
- [12] M. Zhu *et al.*, "Class weights random forest algorithm for processing class imbalanced medical data," *IEEE Access*, vol. 6, 2018.
- [13] H. Boström, "Estimating class probabilities in random forests," in *Proceedings - 6th International Conference on Machine Learning and Applications, ICMLA 2007*, 2007.
- [14] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, 1986.
- [15] L. A. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, 1965.
- [16] W. W. Tan and T. W. Chua, "Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions (Mendel, J.M.; 2001) [book review]," *IEEE Comput. Intell. Mag.*, vol. 2, no. 1, 2007.
- [17] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, 2009.
- [18] D. Patiño-Pérez, L. Armijos-Valarezo, L. Chóez-Acosta, and F. Burgos-Robalino, "Convolutional neural networks for diabetic retinopathy detection," *Ingenius*, no. 33, 2025.
- [19] A. G. Singal *et al.*, "Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma," *Am. J. Gastroenterol.*, vol. 108, no. 11, 2013.
- [20] J. Franklin, "The elements of statistical learning: data mining, inference and prediction," *Mathematical Intelligencer*, vol. 27, no. 2, 2005.
- [21] H. Guo and H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation," *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, 2004.
- [22] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, 2006.
- [23] P. S. Kamath *et al.*, "A model to predict survival in patients with end-stage liver disease," *Gastroenterology*, vol. 120, no. 5, 2001.
- [24] G. D'Amico, G. Garcia-Tsao, and L. Pagliaro, "Natural history and prognostic indicators of survival in cirrhosis: A systematic review of 118

- studies,” *Journal of Hepatology*, vol. 44, no. 1, 2006.
- [25] A. Tripodi and P. M. Mannucci, “The Coagulopathy of Chronic Liver Disease,” *N. Engl. J. Med.*, vol. 365, no. 2, 2011.
 - [26] C. Lescure *et al.*, “Albi score is a strong predictor of toxicity following sirt for hepatocellular carcinoma,” *Cancers (Basel)*, vol. 13, no. 15, 2021.
 - [27] R. B. Freeman *et al.*, “The new liver allocation system: Moving toward evidence-based transplantation policy,” *Liver Transplant.*, vol. 8, no. 9, 2002.
 - [28] G. Riccardo, M. Anna, R. Salvatore, T. Franco, G. Fosca, and P. Dino, “A Survey Of Methods For Explaining Black Box Models,” *ACM Comput. Surv.*, vol. 51, no. 5, 2018.
 - [29] K. G. M. Moons *et al.*, “Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration,” *Ann. Intern. Med.*, vol. 162, no. 1, 2015.
 - [30] S. Wang, Z. Wang, Z. Liu, and J. Wu, “Prognostic value of four immune-related genes in lower-grade gliomas: a biomarker discovery study,” *Front. Genet.*, vol. 15, 2024.
 - [31] T. Neeman, “Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating by Ewout W. Steyerberg,” *Int. Stat. Rev.*, vol. 77, no. 2, 2009.
 - [32] J. M. Mendel, “Type-1 Fuzzy Sets and Fuzzy Logic BT - Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions, 2nd Edition,” in *Uncertain Rule-Based Fuzzy Systems*, 2017.
 - [33] J. S. R. Jang, “ANFIS: Adaptive-Network-Based Fuzzy Inference System,” *IEEE Trans. Syst. Man Cybern.*, vol. 23, no. 3, 1993.
 - [34] M. Tec, “ROC Analysis for Classification and Prediction in Practice,” *J. Am. Stat. Assoc.*, vol. 120, no. 549, 2025.
 - [35] H. Goldstein, P. J. Diggle, K.-Y. Liang, and S. L. Zeger, “Analysis of Longitudinal Data,” *J. R. Stat. Soc. Ser. A (Statistics Soc.)*, vol. 158, no. 2, 1995.
 - [36] A. Spann *et al.*, “Applying Machine Learning in Liver Disease and Transplantation: A Comprehensive Review,” *Hepatology*, vol. 71, no. 3, 2020.