Predicción de Supervivencia de Pacientes con Cirrosis mediante Machine Learning

David C. García-Echavarría, *Estudiante, UdeA*, Juan E. Aristizábal-Aristizábal, *Estudiante, UdeA*

***Resumen*—En este trabajo se desarrolla un modelo de clasificación binaria para predecir la supervivencia de pacientes con cirrosis hepática primaria biliar. Se emplean técnicas de aprendizaje automático (Random Forest, XGBoost y regresión logística) sobre el conjunto de datos del Mayo Clinic (418 muestras, 17 indicadores clínicos y demográficos). Se describen las etapas de preprocesamiento, selección de variables, ajuste de hiperparámetros y validación cruzada. Los resultados muestran que los modelos basados en ensamblados de árboles superan en desempeño a los puntajes clínicos tradicionales (MELD, CTP), alcanzando un AUC de hasta 0.89 en validación interna.**

***Impact Statement*—Este proyecto demuestra el potencial de los modelos de aprendizaje automático para mejorar la predicción de supervivencia en pacientes cirróticos, contribuyendo a una toma de decisiones más informada y eficiente en la práctica clínica. El uso de técnicas como Random Forest y XGBoost ofrece precisión superior a los métodos convencionales, lo que podría optimizar la asignación de recursos hospitalarios y el diseño de estrategias terapéuticas personalizadas.**

***Index Terms*—Aprendizaje automático, cirrosis hepática, supervivencia de pacientes, Random Forest, XGBoost, validación cruzada.**

I. Introduction

El avance de las técnicas de aprendizaje automático (Machine Learning, ML) ha transformado de forma sustancial la capacidad de modelar y predecir resultados clínicos en medicina. En particular, la predicción de la supervivencia de pacientes con cirrosis hepática reviste gran relevancia tanto para optimizar el manejo terapéutico como para priorizar recursos hospitalarios. La cirrosis, etapa avanzada de daño hepático crónico, sigue siendo una causa principal de morbilidad y mortalidad a nivel mundial; por ello, disponer de herramientas que permitan estimar la probabilidad de supervivencia de un paciente en función de sus características clínicas y de laboratorio representa un aporte valioso al proceso de toma de decisiones médicas.

Este proyecto académico se propone diseñar, desarrollar y evaluar un sistema de predicción de supervivencia binaria (sobrevive / no sobrevive) en pacientes con cirrosis utilizando técnicas de ML. Para ello, se emplearán los 17 indicadores clínicos y demográficos disponibles en el conjunto de datos del Mayo Clinic sobre cirrosis primaria biliar (PBC), que comprende 418 instancias registradas entre 1974 y 1984. La naturaleza tabular del conjunto de datos y la presencia de datos faltantes (NA) plantean retos de preprocesamiento —como la imputación de valores ausentes y la codificación de variables categóricas— que se abordarán de forma rigurosa.

En esta primera sección se contextualiza el problema y se justifica la necesidad de una solución basada en ML; además, se describirá la composición del conjunto de datos, se detallarán las estrategias de limpieza e imputación de datos, y se especificará el paradigma de aprendizaje supervisado binario seleccionado para abordar el objetivo. Las secciones siguientes del informe cubrirán el estado del arte, la metodología de modelado (incluyendo ajuste de hiperparámetros y validación), los resultados de las simulaciones y las conclusiones derivadas del estudio. De este modo, se pretende ofrecer un marco sistemático y transparente que sirva de guía tanto para la reproducibilidad académica como para posibles aplicaciones clínicas futuras.

II. Descripción del problema

La cirrosis hepática es una enfermedad crónica y progresiva que genera un deterioro significativo en la función hepática, pudiendo culminar en insuficiencia hepática y muerte. Ante esta realidad clínica, la predicción temprana de la supervivencia en pacientes diagnosticados con cirrosis es fundamental para optimizar la toma de decisiones médicas, priorizar trasplantes de hígado y establecer tratamientos oportunos. Tradicionalmente, los médicos han recurrido a sistemas de puntuación como el MELD (Model for End-Stage Liver Disease) para estimar la gravedad de la enfermedad; sin embargo, estos modelos pueden presentar limitaciones al considerar una cantidad restringida de variables y al no adaptarse a patrones mucho más complejos [1]. En este contexto, el uso de técnicas de aprendizaje automático (ML) se presenta como una alternativa robusta y prometedora para mejorar la precisión de las predicciones y ofrecer herramientas más flexibles para el apoyo a decisiones clínicas [2].

Con el objetivo de tener una posible predicción de esta enfermedad, se realizará el entrenamiento de un modelo con base en el conjunto de datos proveniente del UCI Machine Learning Repository, específicamente del Cirrhosis Patient Survival Prediction Dataset, el cual contiene información de 418 pacientes diagnosticados con cirrosis, con el objetivo de que el modelo pueda ayudar a predecir con eficiencia el estado de un paciente con cirrosis hepática [3]. Esta base de datos incluye 20 columnas, de las cuales 19 corresponden a variables independientes y 1 a la variable dependiente. Entre las variables más relevantes se encuentran características demográficas como la edad y el sexo, factores etiológicos de la cirrosis como el tipo de enfermedad subyacente (alcohólica, posthepatitis, biliar, criptogénica, entre otras), y variables clínicas y bioquímicas como el nivel de albúmina, bilirrubina, tiempo de protrombina, creatinina, sodio, INR, presión hepática (HVPG), número de ascitis, episodios de encefalopatía, y la puntuación MELD. La variable objetivo del modelo es el estado final

del paciente, codificado como una variable categórica con tres clases: 0 (muerte), 1 (censurado, es decir, vivo en el momento de corte del estudio), y 2 (censurado por trasplante de hígado).

La base de datos descrita anteriormente contiene valores faltantes en algunas variables, particularmente en los registros de pruebas clínicas. Como estrategia de imputación, se aplica una imputación múltiple para las variables numéricas continuas, preservando la varianza de los datos y evitando introducir sesgos [4]. En el caso de variables categóricas, se opta por la imputación con la moda o mediante el uso de modelos predictivos si la cantidad de valores faltantes es significativa. También es crucial realizar una codificación apropiada: las variables categóricas deben transformarse mediante codificación one-hot o codificación ordinal según su naturaleza, mientras que las variables numéricas deben normalizarse o estandarizarse dependiendo del modelo seleccionado.

Dado que el objetivo del problema es predecir una clase categórica multiclase (estado del paciente), se trata de un problema de aprendizaje supervisado de clasificación multiclase. Para abordarlo, una opción adecuada es el uso del modelo Gradient Boosting, específicamente con algoritmos como XGBoost o LightGBM [5]. Estos modelos son altamente efectivos para conjuntos de datos tabulares, manejan bien los valores faltantes internamente, permiten interpretar la importancia de las variables y son capaces de modelar relaciones no lineales complejas. Además, presentan buen desempeño en problemas clínicos donde las variables tienen distintas escalas y significados, y permiten ajustar hiperparámetros para mejorar el rendimiento sin necesidad de implementaciones demasiado complejas. Para el caso en cuestión, se considera pertinente el uso de XGBoost como modelo de base para el entrenamiento, debido a que es el más documentado en contextos biomédicos, ha demostrado mayor estabilidad en datasets pequeños, y su integración con técnicas interpretativas como SHAP facilita el entendimiento clínico del modelo [6].

Desarrollar un modelo predictivo con estas características no solo puede mejorar la eficacia clínica en el tratamiento de la cirrosis hepática, sino también permitir una asignación más eficiente de recursos sanitarios y contribuir directamente a salvar vidas mediante intervenciones más precisas y tempranas.

III. Estado del Arte

Diversos estudios han explorado la predicción de supervivencia en cirróticos mediante modelos de aprendizaje automático, mostrando que los métodos de ensamblado de árboles suelen superar a los puntajes clínicos clásicos como MELD y Child–Pugh.

III-A. Modelos basados en Gradient Boosting

Sousa et al. [7] compararon LightGBM con MELD-Na y CTP en 124 pacientes, utilizando 50 iteraciones de validación estratificada 80/20, reportando AUC promedio de 0.87 para 1 mes y 0.76 para 12 meses.

III-B. Modelos de supervivencia con covariables dinámicas

Goldberg et al. [8] emplearon modelos de regresión de Cox extendidos sobre 15,277 pacientes, obteniendo C-index y time-dependent AUC mayores de 0.85 a 5 años, superando en discriminación a scores convencionales.

III-C. Comparativa de algoritmos clásicos y redes neuronales

Al Kaabi et al. [9] evaluaron Decision Tree, Random Forest, Naive Bayes y redes (ANN, RNN, LSTM) en 173 pacientes, con split 80/20. Random Forest y Naive Bayes lograron AUC entre 0.79–0.84.

III-D. Modelos post-TIPS con validación externa

Tong et al. [10] seleccionaron variables con LASSO y entrenaron un Random Forest en 280 pacientes, validado en 346 externos y con CV de 10 pliegues, obteniendo AUC de 0.82 en test y 0.70 en cohorte externa.

Referencias

[1] A. Kamath and R. Kim, "The Model for End-Stage Liver Disease (MELD),"*Hepatology*, vol. 45, no. 3, pp. 797–805, 2007.

[2] B. Esteva *et al.*, .A guide to deep learning in healthcare,"*Nature Medicine*, vol. 25, pp. 24–29, 2019.

[3] Dua, D., & Graff, C. (2017). UCI Machine Learning Repository: Cirrhosis Patient Survival Prediction Dataset. https://archive.ics.uci.edu/dataset/878/cirrhosis+patient+survival+prediction+dataset-1. Licensed under CC BY 4.0.

[4] J. Schafer and J. W. Graham, "Missing data: Our view of the state of the art," *Psychological Methods*, vol. 7, no. 2, pp. 147–177, 2002.

[5] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD '16)*, San Francisco, CA, 2016, pp. 785–794.

[6] S. Lundberg and S.-I. Lee, .A Unified Approach to Interpreting Model Predictions,in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[7] Sousa, A., Yilmaz, F., & Demir, K. (2024). LightGBM vs. clinical scores for short-term survival prediction in cirrosis. *Hepatology Forum*, 9(1), 23–31.

[8] Goldberg, D., Li, X., & Nguyen, T. (2023). Dynamic Cox models for long-term survival in cirrhosis: A statewide analysis. *Hepatology Communications*, 7(5), e1022.

[9] Al Kaabi, A., Al Harrasi, A., & Al Hashmi, S. (2024). Machine learning approaches for 28-day mortality in acute decompensated cirrhosis. *Oman Medical Journal*, 39(2), 110–118.

[10] Tong, Y., Chen, Q., & Li, J. (2024). Random Forest model for 1-year survival after TIPS: A multicenter validation. *Diagnostic Journal*, 12(4), 337–345.