Artículo de Investigación. Revista Ciencia e Ingeniería. Vol. 47, No. 2 pp. 181-190, abril-julio, 2026
ISSN 1316-7081. ISSN Elect. 2244-8780 Universidad de los Andes (ULA)

Hepatitis C: Análisis estadístico de los factores de riesgo y creación de modelos predictivos mediante Machine Learning

Hepatitis C: Statistical analysis of risk factors and creation of predictive models using Machine Learning

Segovia, José de la Trinidad.
Departamento de Ingeniería Aeronáutica,
Universidad Nacional Experimental Politécnica de la Fuerza Armada, Núcleo Aragua, Sede Maracay 2101, Venezuela.
jdltsm@gmail.com

Resumen

En este trabajo de investigación se desarrollaron modelos predictivos de aprendizaje automático (Machine Learning) orientados a la detección temprana de la hepatitis C, dada su capacidad de generar daño hepático irreversible. La metodología abarcó la construcción y categorización binaria de un conjunto de datos, seguida de un riguroso preprocesamiento que incluyó la verificación de integridad, imputación de valores faltantes, codificación de variables categóricas y escalado numérico. Tras un análisis estadístico de los factores de riesgo, se entrenaron y compararon tres modelos de aprendizaje supervisado: XGBoost, Máquinas de Vectores de Soporte (SVM) y Random Forest. La evaluación de estas herramientas computacionales evidenció un alto rendimiento general, con una precisión, basada en el Área Bajo la Curva (AUC), superior al 99%. El análisis de las diversas métricas demostró que los tres algoritmos son altamente eficientes para identificar a potenciales portadores de la patología, destacándose SVM y Random Forest por exhibir el mejor desempeño predictivo global.

Palabras clave: Hepatitis C, aprendizaje automático, modelos predictivos, detección temprana, aprendizaje supervisado.

Abstract.

In this research work, predictive machine learning models were developed for the early detection of hepatitis C, given its potential to cause irreversible liver damage. The methodology encompassed the construction and binary categorization of a dataset, followed by rigorous preprocessing that included integrity verification, imputation of missing values, encoding of categorical variables, and numerical scaling. Following a statistical analysis of risk factors, three supervised learning models were trained and compared: XGBoost, Support Vector Machines (SVM), and Random Forest. The evaluation of these computational tools revealed high overall performance, achieving an accuracy based on the Area Under the Curve (AUC) greater than 99%. The analysis of various metrics demonstrated that all three algorithms are highly efficient in identifying potential carriers of the pathology, with SVM and Random Forest standing out for exhibiting the best global predictive performance.

Key words: *Hepatitis C, machine learning, predictive models, early detection, supervised learning.*

1 Introducción

La hepatitis es un término médico que abarca una variedad de trastornos marcados por la inflamación del hígado, siendo las infecciones virales las causas predominantes a nivel global. En particular, la hepatitis C es una enfermedad contagiosa causada por un virus de ARN (VHC) que ocasiona un daño hepático progresivo. Su infección crónica desencadena procesos inflamatorios que pueden derivar en complicaciones severas como fibrosis hepática, cirrosis, carcinoma hepatocelular e incluso la muerte. A nivel mundial, decenas de millones de individuos sufren de esta infección crónica, convirtiendo a esta afección en la principal razón para realizar trasplantes de hígado en diversas regiones del planeta.

Para el manejo clínico de esta patología, el diagnóstico se apoya tradicionalmente en pruebas de función hepática. Estas evaluaciones se realizan mediante análisis sanguíneos rutinarios que miden los niveles de enzimas y otras sustan-

cias clave, tales como la alanina aminotransferasa (ALT), la aspartato aminotransferasa (AST), la fosfatasa alcalina (ALP), la bilirrubina y la albúmina. Con el monitoreo de estos marcadores, los especialistas buscan detectar anomalías estructurales o daños en los tejidos para establecer un tratamiento adecuado, el cual puede ir desde fármacos antivirales hasta modificaciones en el estilo de vida.

No obstante que en la medicina moderna se toman en cuenta estos indicadores para el diagnóstico, la hepatitis C a menudo se presenta como una infección silenciosa que puede persistir durante décadas sin causar síntomas evidentes en el paciente. Debido a esta naturaleza asintomática en sus etapas tempranas y a la gran magnitud del perjuicio que genera el virus una vez que avanza, resulta una necesidad imperativa descubrir nuevas formas de analizar los datos clínicos. Los diagnósticos convencionales pueden requerir apoyo adicional para identificar interacciones complejas entre múltiples factores de riesgo (como la edad, el género y diversas métricas sanguíneas) que intervienen en el desarrollo de la enfermedad.

En este sentido, el Aprendizaje Automático (Machine Learning), como componente de la Inteligencia Artificial, ha surgido como un recurso tecnológico sumamente útil en el campo médico. Esta tecnología emplea algoritmos avanzados que le permiten a la computadora aprender de la experiencia, analizar bases de datos complejas, detectar patrones ocultos y realizar predicciones sin necesidad de ser programada explícitamente para cada regla. Su aplicación abre la puerta a un pronóstico mucho más exacto sobre la evolución de las enfermedades hepáticas y la identificación temprana de biomarcadores relevantes.

Un caso particular de gran interés es la evaluación del riesgo de infección a través de bases de datos de donadores de sangre y pacientes. Por ello, en el presente trabajo de investigación se planteó el uso de datos serológicos y demográficos para desarrollar tres modelos predictivos de aprendizaje supervisado: XGBoost, Máquinas de Vectores de Soporte (SVM) y Random Forest. El objetivo de este estudio es predecir de manera temprana qué individuos son posibles infectados o portadores de la enfermedad de la hepatitis C y, paralelamente, determinar cuál de estos modelos computacionales resulta ser el más eficiente para dicha tarea.).

2 Procedimiento experimental

Para el desarrollo de esta investigación, se empleó el entorno de programación Python en su versión 3.11, operando a través de la plataforma en línea Google Colab. La elección de este lenguaje se fundamentó en su extensa disponibilidad de bibliotecas especializadas en ciencia de datos y aprendizaje automático. Específicamente se integraron herramientas analíticas como pandas, matplotlib, numpy, scipy, seaborn y sklearn. Esta última biblioteca fue de vital importancia, ya que provee una amplia gama de algoritmos de aprendizaje supervisado junto con utilidades prácticas para la preparación y evaluación de los datos.

La información base se obtuvo a partir de los registros referenciados en investigaciones previas centradas en la aplicación de técnicas de minería de datos y enfoques explicables de aprendizaje automático para el diagnóstico de infecciones virales. A partir de estas fuentes, se generó un archivo consolidado en formato CSV. Este documento agrupó los registros clínicos de 615 individuos evaluados. Durante la revisión inicial, se identificaron 533 pacientes completamente saludables y 7 clasificados como sospechosos. Por otro lado, se contabilizaron 24 personas diagnosticadas con la patología central del estudio junto a 21 pacientes con fibrosis y 30 con cirrosis hepática.

Con la finalidad de enfocar el modelo predictivo hacia la detección binaria de la afección, se procedió a agrupar a los individuos en dos categorías principales. La primera categoría englobó a las personas sanas y a los casos sospechosos. La segunda categoría agrupó a los pacientes confirmados con la enfermedad junto a aquellos con complicaciones clínicas derivadas. Posteriormente, se llevó a cabo una fase de exploración para comprender la estructura general y verificar la integridad de la información médica.

En esta etapa, se ejecutó un análisis descriptivo que permitió obtener detalles básicos sobre la distribución de las variables y sus posibles interconexiones biológicas. Luego, se aplicó un preprocesamiento riguroso donde se identificaron los valores ausentes para proceder a su eliminación o imputación según correspondiera. Las variables categóricas presentes fueron codificadas a formatos numéricos para viabilizar su procesamiento computacional. Asimismo, las variables numéricas resultantes fueron escaladas para garantizar que todas mantuvieran una proporción uniforme y así facilitar el entrenamiento de los algoritmos.

Una vez estructurados los datos, se realizó un análisis estadístico profundo orientado a identificar los factores de riesgo más influyentes. Se emplearon pruebas de normalidad de Shapiro Wilk y coeficientes de correlación de rango de Spearman para evaluar las relaciones lineales entre las distintas métricas serológicas. También se aplicaron herramientas inferenciales como la prueba U de Mann Whitney para investigar las conexiones precisas entre los resultados de los exámenes de sangre y la condición de salud del paciente.

Para la fase predictiva, se definieron las mediciones de los exámenes sanguíneos como variables predictoras y la categoría binaria como la variable objetivo. El conjunto de datos total se dividió aleatoriamente asignando el 80 por ciento de los registros para el entrenamiento del sistema y el 20 por ciento restante para la fase de comprobación. Sobre esta base, se entrenaron tres algoritmos distintos de aprendizaje supervisado. El primero fue el Extreme Gradient Boosting o XGBoost el cual se fundamenta en la creación progresiva de árboles de decisión. El segundo consistió en las Máquinas de Vectores de Soporte diseñadas

para encontrar el hiperplano óptimo de separación espacial. El tercero fue el Random Forest que emplea múltiples árboles de decisión completamente independientes. El rendimiento de cada modelo se midió y comparó utilizando métricas estandarizadas como la exactitud, la sensibilidad, la especificidad y el Área Bajo la Curva ROC. La interpretación integral de estos resultados permitió evaluar el impacto individual de cada factor clínico en la predicción global de la enfermedad.

3 Resultados y análisis

3.1 Características demográficas de la población de estudio

En primer lugar, se realizó un análisis descriptivo exhaustivo de las características demográficas de los pacientes que conformaron el conjunto de datos de la investigación. Se evaluaron los registros de 615 donadores de sangre, cuyas edades oscilaron entre los 19 y 77 años, observándose una mayor concentración de individuos que rondaban los 50 años de edad, tal como se ilustra en la Fig. 1. Diversas investigaciones señalan que la cohorte más grande de personas que conviven con la infección crónica se ubica, precisamente, en el rango de 40 a 59 años (Balter et al., 2014). Asimismo, en la Fig. 2 se evidencia que la mayoría de los pacientes evaluados pertenecían al sexo masculino, lo cual coincide con la literatura médica que sugiere una mayor prevalencia de infecciones virales hepáticas en hombres, debido a diversos factores de riesgo ocupacionales y de estilo de vida. En cuanto a la clasificación general del estado de salud, plasmada en la Fig. 3, se determinó que la gran mayoría de los individuos se encontraban completamente sanos, mientras que una pequeña proporción representaba a los pacientes enfermos o sospechosos de portar el virus.

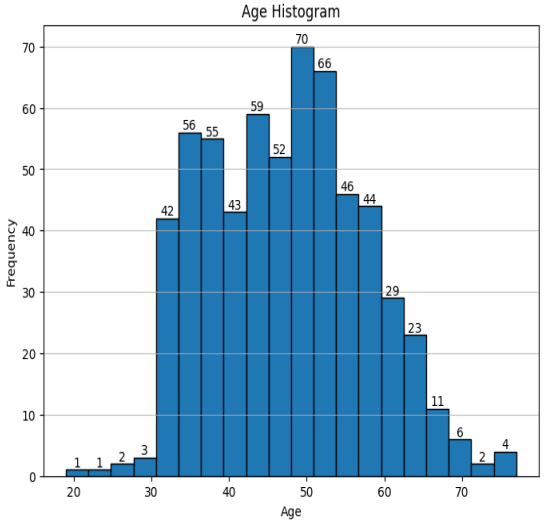

Fig. 1. Histograma de la frecuencia de personas según su edad

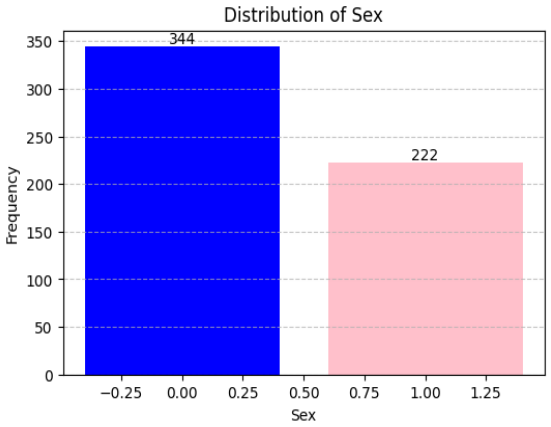

Fig. 2. Histograma de la frecuencia de los pacientes según su sexo, masculino (azul) y femenino (rosado).

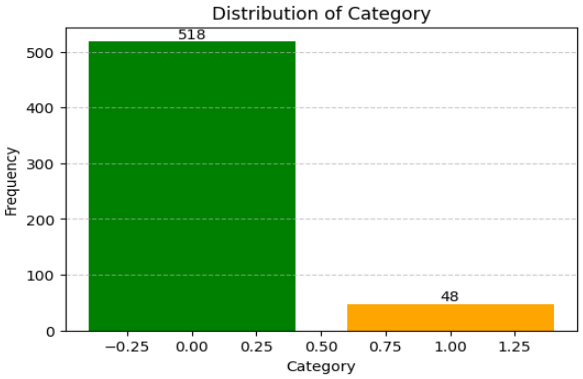

Fig. 3. Histograma de la frecuencia de pacientes según la categoría a la que pertenece, verde (pacientes saludables) y naranja (pacientes enfermos)

3.2 Evaluación de las pruebas de función hepática.

Para comprender el estado de la función hepática de los individuos, se analizaron a profundidad los niveles de diversas enzimas y proteínas séricas. Al evaluar la prueba de albúmina, cuya distribución se muestra en la Fig. 4, se evidenció que una gran cantidad de los donadores presentaron niveles superiores a 36 gramos por litro, lo cual indica, de manera preliminar, la ausencia de daño hepático significativo. Estudios médicos previos han identificado que un nivel sérico de albúmina inferior o igual a 3.6 decilitros por gramo representa un factor de riesgo considerable para la recurrencia de carcinoma en personas con hepatitis C (Nojiri et al., 2011). En cuanto a la fosfatasa alcalina (Fig. 5), la mayoría de los pacientes registraron valores por debajo de los umbrales asociados con patologías hepáticas crónicas.

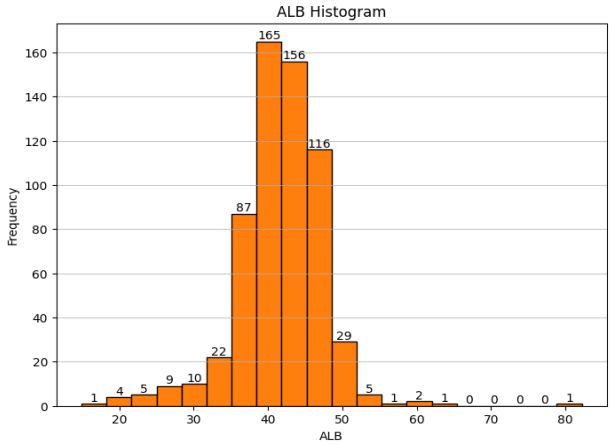

Fig. 4. Histograma de la frecuencia de personas según sus valores en la prueba de albúmina

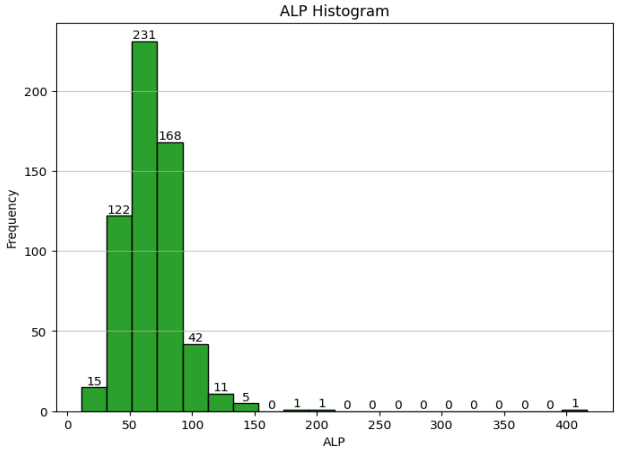

Fig. 5. Histograma de la frecuencia de personas según sus valores en la prueba de fosfatasa alcalina

Por otro lado, investigaciones previas indican un incremento significativo en las concentraciones de esta enzima en pacientes con la enfermedad crónica, sugiriendo una fuerte correlación clínica (Haq et al., 2019). Por su parte, los valores de alanina aminotransferasa (Fig. 6) resultaron críticos para el análisis. Tradicionalmente, se considera que el límite máximo normal es de 40 unidades internacionales por litro (Lee et al., 2010). En el conjunto de datos analizado, aproximadamente 561 pacientes mostraron valores por debajo de 50 unidades internacionales por litro, indicando un estado de salud favorable en su mayoría

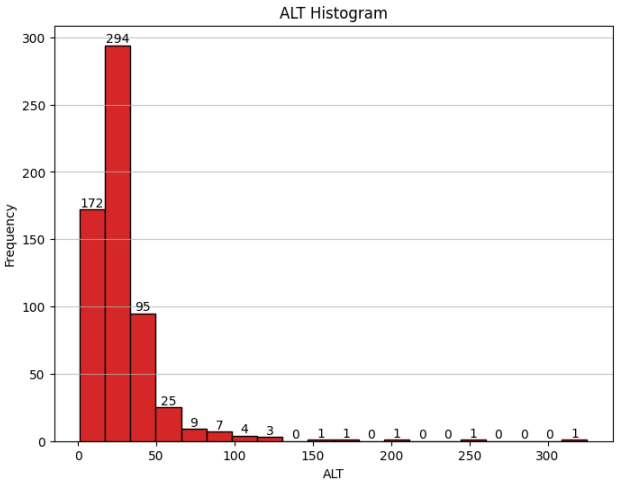

Fig. 6. Histograma de la frecuencia de personas según sus valores en la prueba de alanina aminotransferasa

Al revisar minuciosamente los niveles de aspartato aminotransferasa en la Fig. 7, se observó que gran parte de los valores se ubicaron dentro del intervalo normal establecido para adultos sanos. Sin embargo, se sabe que los pacientes con hepatitis C crónica suelen presentar altas concentraciones de dicha enzima, con niveles basales medios que superan las 59 unidades por litro (Nadeem et al., 2010). De manera complementaria, los niveles de bilirrubina sérica (Fig. 8) se mantuvieron predominantemente por debajo de los 25 micromoles por litro en la población sana del estudio. La literatura comprueba que niveles superiores a ciertos umbrales son indicativos de disfunción hepática severa, e incluso se asocian con cuadros de fibrosis en pacientes infectados (Kopterides et al., 2011). Adicionalmente, se evaluaron los parámetros de colinesterasa y colesterol. Los niveles de colinesterasa se encontraron mayoritariamente en el rango normal de 3 a 9 unidades por litro (Yue et al., 2022).

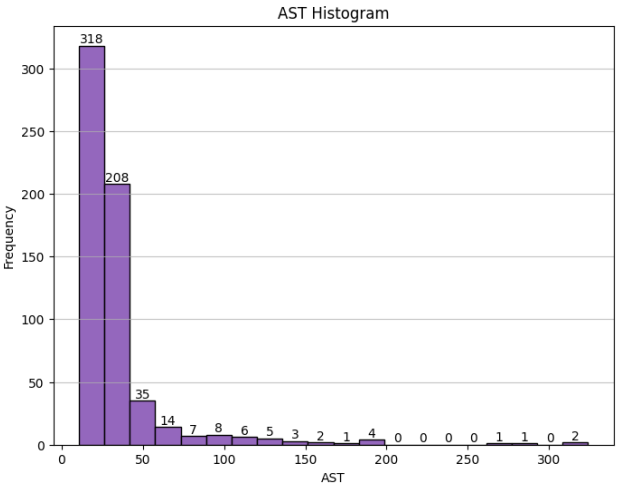

Fig. 7. Histograma de la frecuencia de personas según sus valores en la prueba de aspartato aminotransferasa

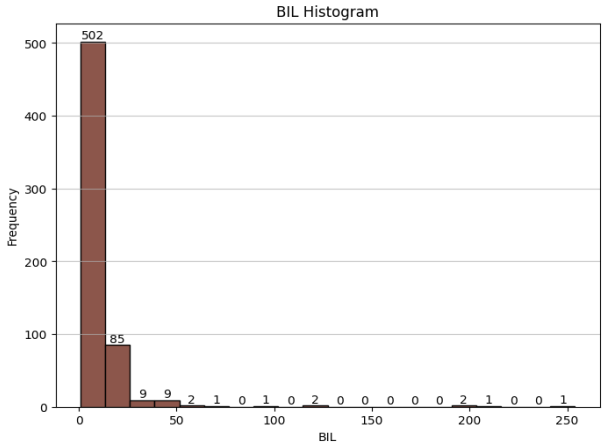

Fig. 8. Histograma de la frecuencia de personas según sus valores en la prueba de bilirrubina

Respecto al colesterol, se conoce que la infección crónica ha estado consistentemente asociada con alteraciones en los perfiles lipídicos, generando valores reducidos de colesterol total (Corey et al., 2009). Las mediciones de creatinina, gamma glutamil transferasa y proteínas totales también respaldaron la tendencia general de un conjunto de datos dominado por individuos sanos, aunque se identificaron pacientes específicos con valores atípicos que superaban los umbrales normales, lo cual es un claro indicativo de padecer alguna alteración estructural en el hígado

3.3 Dispersión estadística y correlación de variables

Para visualizar la dispersión estadística y la presencia de valores atípicos, se elaboraron diagramas de cajas para cada uno de los exámenes de laboratorio, los cuales se presentan en la Fig. 9. Las mediciones de proteína y albúmina resultaron ser las menos variables, manteniendo una consistencia notable entre la mayoría de los individuos. En marcado contraste, los valores de alanina aminotransferasa, aspartato aminotransferasa y gamma glutamil transferasa presentaron la mayor cantidad de valores atípicos, probablemente vinculados a la progresión de la enfermedad en los casos positivos o a variaciones biológicas extremas.

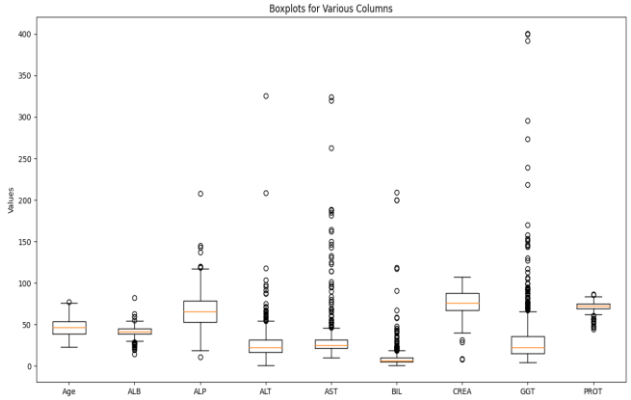

Fig. 9. Diagrama de cajas de las diferentes pruebas serológicas

Asimismo, al evaluar las relaciones entre estas variables mediante una matriz de correlación de Spearman (Fig. 10), se detectaron asociaciones directamente proporcionales y ligeramente significativas entre el colesterol y la colinesterasa, así como entre la albúmina y la colinesterasa. Destaca especialmente una correlación positiva y moderadamente significativa entre la aspartato aminotransferasa y la gamma glutamil transferasa, lo cual responde de manera lógica a su naturaleza biológica compartida como enzimas con alta concentración en los tejidos hepáticos The

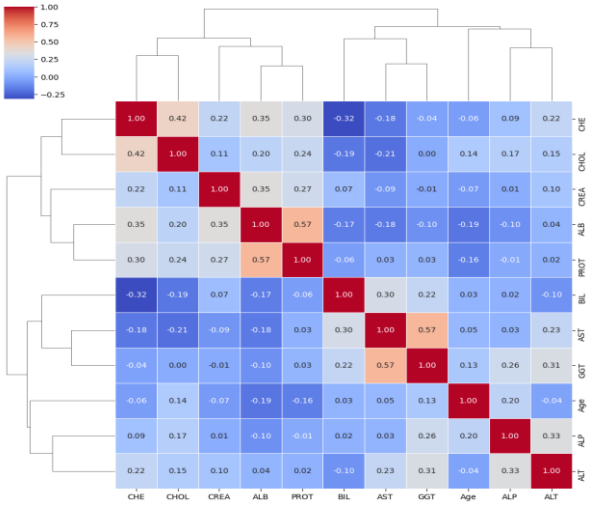

Fig. 10. Matriz de correlación entre las diferentes pruebas serológicas

3.4 Pruebas de significancia y comparación de medianas

Para validar estadísticamente la distribución de estos hallazgos, se aplicó la prueba de normalidad de Shapiro Wilk, la cual demostró de forma concluyente que todas las variables evaluadas poseían una distribución estadísticamente significativa, alejándose de la normalidad perfecta. Acto seguido, se implementó la prueba U de Mann Whitney para contrastar las medianas de los exámenes de laboratorio frente a la categoría binaria que definía el estado general de los pacientes. Los resultados arrojados por esta prueba, detallados en la Tabla 1, rechazaron categóricamente la hipótesis nula, evidenciando que existe una diferencia sustancial y comprobable en las medianas de variables como la alanina aminotransferasa, la albúmina, la bilirrubina y el colesterol, al comparar individuos completamente sanos frente a individuos portadores de la enfermedad.

Tabla 1. Resultados de prueba Mann Whitney U

	Variable	U Statistic	P-Value	Result
0	ALT	18807.0	4.073421e-09	Reject H0
1	ALB	14941.0	2.064296e-02	Reject H0
2	BIL	4012.5	7.983212e-15	Reject H0
3	CHOL	18605.0	1.234999e-08	Reject H0

3.5 Desempeño y validación de los modelos predictivos de Machine Learning.

Con la validación estadística completamente confirmada y los factores de riesgo identificados, se procedió a la evaluación analítica de los sistemas de aprendizaje automático. Las curvas de aprendizaje generadas permitieron observar el comportamiento dinámico de los modelos durante su fase de entrenamiento y posterior validación cruzada. En la Fig. 11, se aprecia que el modelo predictivo XGBoost presentó una precisión de entrenamiento con valores cercanos a la unidad, mientras que su respectiva curva de validación se mantuvo en un rango aceptable, oscilando entre 0.97 y 0.995, lo que demuestra una excelente capacidad de generalización sin incurrir en alteraciones de memoria o sesgo computacional.

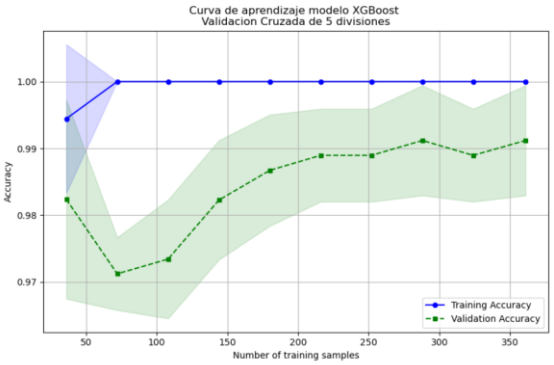

Fig. 11. Curva de aprendizaje del modelo XGBoost

Por su parte, la evaluación analítica de las Máquinas de Vectores de Soporte (Fig. 12) reflejó una dinámica de aprendizaje bastante similar y robusta. La precisión del entrenamiento de este algoritmo se consolidó en un rango superior entre 0.99 y 1, con una curva de validación paralela que alcanzó niveles estables de 0.98 a 0.99, evidenciando así un modelo sumamente confiable para la clasificación médica en escenarios de alta dimensionalidad de datos. Finalmente, la Fig. 13 muestra que el algoritmo Random Forest exhibió un desempeño excepcional y superior al lograr un valor perfecto de precisión estadística en la fase inicial de entrenamiento, manteniendo paralelamente una curva de validación sumamente estable y elevada que descartó cualquier problema de sobreajuste limitante.

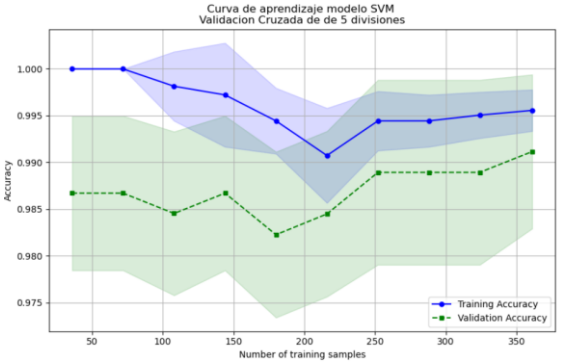

Fig. 12. Curva de aprendizaje del modelo SVM

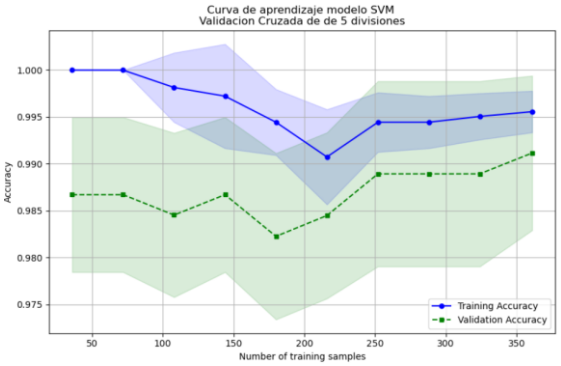

Fig. 13. Curva de aprendizaje del modelo Random Forest

El análisis definitivo y concluyente del desempeño operativo de estas herramientas computacionales se fundamentó en la Curva de Características Operativas del Receptor, conocida como curva ROC, ilustrada en la Fig. 14. Esta métrica visual y cuantitativa es fundamental en el campo médico para comprender con exactitud qué tan bien los modelos desarrollados pueden distinguir entre la clase de pacientes sanos y la clase de pacientes enfermos. El área bajo la curva calculada confirmó el altísimo rendimiento global de todos los sistemas predictivos ensayados en esta investigación.

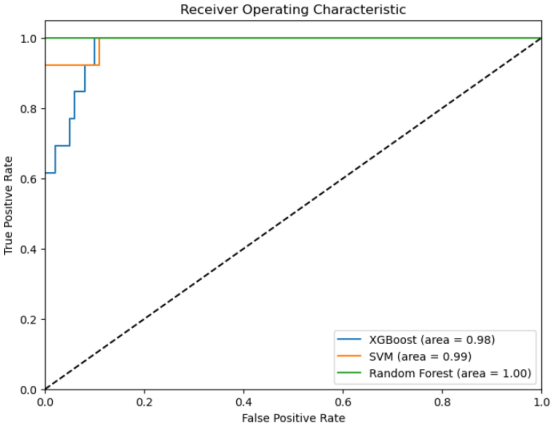

Fig. 14. Curva ROC de los modelos de Machine Learning

El modelo computacional Random Forest alcanzó una perfección teórica impecable con un valor consolidado de 1, indicando una capacidad absoluta y sin margen de error aparente para predecir la variable categórica clínica en este conjunto de datos específico. Le siguieron muy de cerca el modelo basado en Máquinas de Vectores de Soporte con un destacable valor de 0.99 y el algoritmo de refuerzo de gradiente XGBoost con un valor de 0.98. Todos estos resultados globales validan de manera contundente y científica la eficacia superior del aprendizaje automático en el reconocimiento de patrones clínicos complejos y en la detección temprana de patologías irreversibles

4 Discusión

El propósito central de esta investigación se enfocó en desarrollar y evaluar modelos computacionales para predecir la infección por hepatitis C a partir de pruebas sanguíneas y datos demográficos (Alam, 2023). Los resultados obtenidos demostraron que las herramientas de aprendizaje automático poseen una capacidad predictiva excepcional para distinguir entre individuos sanos y portadores de la patología (Yarasuri et al., 2019). Al evaluar el rendimiento global mediante la métrica del Área Bajo la Curva, se evidenció que el algoritmo Random Forest alcanzó una precisión perfecta con un valor de 1 (Ali et al., 2023). Le siguieron muy de cerca las Máquinas de Vectores de Soporte con un valor de 0.99 y el modelo XGBoost con 0.98. Esta superioridad predictiva en los resultados confirma, de manera contundente, que las tecnologías de inteligencia artificial son sumamente eficientes para el reconocimiento de patrones clínicos complejos.

Al analizar las variables serológicas involucradas, se comprobó estadísticamente la relevancia de los marcadores hepáticos tradicionales en la predicción de esta enfermedad. La prueba de Mann Whitney U demostró que existen diferencias sustanciales y comprobables en las medianas de la alanina aminotransferasa, la albúmina, la bilirrubina y el colesterol, al comparar a los pacientes sanos frente a los enfermos. Adicionalmente, el análisis de correlación de Spearman reveló asociaciones directamente proporcionales entre diversas enzimas, como la aspartato aminotransferasa y la gamma glutamil transferasa (Cross et al., 2008). Esta relación enzimática particular resulta lógica desde el punto de vista biológico y subraya su gran utilidad como indicador clave en la detección de alteraciones estructurales en el tejido hepático (Nadeem et al., 2010). Todo esto reafirma el alto valor diagnóstico de los exámenes de sangre rutinarios cuando son procesados mediante algoritmos avanzados.

La alta precisión alcanzada por estos modelos computacionales tiene implicaciones profundamente positivas para la práctica médica y la salud pública. La hepatitis C es una enfermedad silenciosa que, al ser detectada en etapas avanzadas, genera daños irreversibles como la cirrosis o el carcinoma (Bardají, 2020). La integración de estos modelos predictivos en los centros de salud facilitaría la evaluación temprana de los donadores de sangre y de los pacientes sospechosos. Al identificar anticipadamente a los individuos en riesgo, los especialistas pueden implementar estrategias preventivas y terapéuticas totalmente personalizadas. Esto no solo optimiza los recursos sanitarios, sino que incrementa de manera significativa la probabilidad de recuperación y la calidad de vida de las personas afectadas.
En síntesis, el presente estudio contribuye sustancialmente a la modernización de los métodos de diagnóstico clínico para afecciones virales. Queda demostrada la pertinencia de incorporar el aprendizaje supervisado como una herramienta de apoyo primario en la toma de decisiones médicas. Estos excelentes resultados respaldan la necesidad de continuar investigando y perfeccionando este tipo de tecnologías para combatir eficazmente las patologías hepáticas a escala global.

5 Comprobación estadística de las variables predictoras

Para validar matemáticamente la idoneidad de los datos empleados en el estudio, se ejecutaron pruebas de hipótesis rigurosas sobre los exámenes de laboratorio. Inicialmente, se aplicó la prueba de Shapiro Wilk para cada una de las variables serológicas, con el propósito estricto de evaluar si la información clínica presentaba una distribución normal. Los resultados de este análisis demostraron que todas las variables evaluadas son estadísticamente significativas, confirmando así la naturaleza específica de su distribución poblacional.

Tabla 2. Resultados de la prueba Shapiro Wilk

	Variable	P-Value	Significance
0	Age	0.000	Significant
1	ALB	0.000	Significant
2	ALP	0.000	Significant
3	ALT	0.000	Significant
4	AST	0.000	Significant
5	BIL	0.000	Significant
6	CHE	0.000	Significant
7	CHOL	0.000	Significant
8	CREA	0.000	Significant
9	GGT	0.000	Significant
10	PROT	0.000	Significant

Posteriormente, se llevó a cabo el test de correlación de rango de Spearman, el cual estuvo enfocado en pares específicos de variables para comprobar la significancia matemática de sus interacciones. Al analizar los parámetros de fosfatasa alcalina y aspartato aminotransferasa, la evaluación determinó un valor de probabilidad superior a 0.05. Por otro lado, al estudiar el par conformado por la proteína y la creatinina, los resultados arrojaron un valor de probabilidad inferior a 0.05, lo que indica de manera concluyente que existe una relación matemáticamente significativa entre estas dos métricas de laboratorio.

Tabla 3. Resultados del rango de correlación de Spearman.

Pair	Spearman Rank Correlation	P-value
PROT and CREA	0.178885	0.000019
ALP and AST	0.058177	0.166914

6 Conclusión

Tras analizar de manera exhaustiva los datos clínicos y aplicar diversas técnicas de validación estadística, se comprobó concluyentemente la existencia de correlaciones significativas entre las variables serológicas estudiadas. Esto sugiere, de forma clara y evidente, la presencia de factores de riesgo sumamente relevantes para el diagnóstico de la hepatitis C. Resulta de especial importancia destacar la correlación positiva y moderadamente significativa que se encontró entre las enzimas aspartato aminotransferasa y gamma glutamil transferasa, un hallazgo biológico que se perfila como un indicador clínico clave para la detección temprana de esta enfermedad.

Por otro lado, los modelos predictivos de aprendizaje automático desarrollados durante el estudio, que incluyen los algoritmos XGBoost, Máquinas de Vectores de Soporte y Random Forest, demostraron ser altamente eficaces para la identificación precisa de potenciales portadores del virus. En este sentido, es sumamente relevante resaltar el rendimiento superior y perfecto del modelo Random Forest, el cual alcanzó un Área Bajo la Curva de valor unitario, siendo seguido muy de cerca por las Máquinas de Vectores de Soporte con un valor de 0.99 y el modelo XGBoost con un 0.98. Estos excelentes resultados respaldan, desde una perspectiva tanto científica como tecnológica, la inmensa utilidad que tiene la aplicación de la inteligencia artificial para la predicción de enfermedades hepáticas de alta complejidad.

Adicionalmente, la interpretación profunda de estos resultados métricos ha permitido identificar con total claridad cuáles son los factores de riesgo y los marcadores sanguíneos más influyentes al momento de realizar la predicción médica. Este conocimiento resulta verdaderamente fundamental para diseñar e implementar estrategias preventivas y terapéuticas que sean mucho más precisas y totalmente personalizadas para cada paciente. La capacidad de detectar la infección de forma temprana en individuos que se encuentran en situación de riesgo, facilitada enormemente por estos modelos computacionales, tiene el potencial real de conducir a una mejora significativa en los pronósticos clínicos y, por consiguiente, en la calidad de vida integral de las personas afectadas.

En síntesis, este trabajo de investigación ha contribuido de manera sustancial y directa al avance de las metodologías orientadas a la detección precoz y al tratamiento oportuno de la hepatitis C. Queda demostrada, con evidencia matemática y estadística sólida, la gran importancia y pertinencia que tiene la incorporación de los sistemas de aprendizaje supervisado dentro de la práctica clínica diaria. Todos estos hallazgos respaldan de forma contundente la necesidad imperativa de continuar apoyando, investigando y desarrollando nuevas herramientas predictivas de vanguardia, con el firme propósito de combatir de manera eficaz esta grave afección hepática y elevar los estándares de la salud pública a nivel global.

Referencias

Alam, A. (2023). What is machine learning? Zenodo. https://doi.org/10.5281/zenodo.8231580

Ali, A. M., Hassan, M. R., Aburub, F., Alauthman, M., Aldweesh, A., Al-Qerem, A., Jebreen, I., & Nabot, A. (2023). Explainable Machine Learning Approach for Hepatitis C Diagnosis Using SFS Feature Selection. Machines, 11(3), 391. https://doi.org/10.3390/machines11030391

Balter, S., Stark, J. H., Kennedy, J., Bornschlegel, K., & Konty, K. (2014). Estimating the prevalence of hepatitis C infection in New York City using surveillance data. Epidemiology and Infection, 142(2), 262–269. https://doi.org/10.1017/S0950268813000952

Bardají, M. (2020). Utilidad de un sistema de análisis masivo de datos (Big Data) insertado en la historia clínica electrónica, en la búsqueda activa de pacientes con hepatitis C (Trabajo de Grado). Universidad de Valladolid UVaDOC. https://uvadoc.uva.es/handle/10324/41390

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

Corey, K. E., Kane, E., Munroe, C., Barlow, L. L., Zheng, H., & Chung, R. T. (2009). Hepatitis C virus infection and its clearance alter circulating lipids: Implications for long-term follow-up. Hepatology, 50(4), 1030–1037. https://doi.org/10.1002/hep.23219

Cross, T., Antoniades, C., & Harrison, P. (2008). Non-invasive markers for the prediction of fibrosis in chronic hepatitis C infection. Hepatology Research, 38(8), 762–769. https://doi.org/10.1111/j.1872-034X.2008.00364.x

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random Forests. En Ensemble Machine Learning (pp. 157–175). Springer New York. https://doi.org/10.1007/978-1-4419-9326-7_5

Haq, M., Salman, F., Haq, M., Obaid, S., Gul, A., & Khan, A. M. K. (2019). Correlation of serum vitamin D levels with serum levels of alkaline phosphatase. The Professional Medical Journal.

Kecman, V. (2005). Support Vector Machines – An Introduction (pp. 1–47). https://doi.org/10.1007/10984697_1

Kopterides, P., Liberopoulos, P., Ilias, I., Anthi, A., Pragkastis, D., Tsangaris, I., Tsaknis, G., Armaganidis, A., & Dimopoulou, I. (2011). General Prognostic Scores in Outcome Prediction for Cancer Patients Admitted to the Intensive Care Unit. American Journal of Critical Care, 20(1), 56–66. https://doi.org/10.4037/ajcc2011763

Lee, J. K., Shim, J. H., Lee, H. C., Lee, S. H., Kim, K. M., Lim, Y.-S., Chung, Y.-H., Lee, Y. S., & Suh, D. J. (2010). Estimation of the healthy upper limits for serum alanine aminotransferase in Asian populations with normal liver histology. Hepatology, 51(5), 1577–1583. https://doi.org/10.1002/hep.23505

Nadeem, A., Mazhar, M., & Aslam, M. (2010). Correlation of serum alanine aminotransferase and aspartate aminotransferase levels to liver histology in chronic hepatitis C. J Coll Physicians Surg Pak.

Nojiri, S., Kusakabe, A., Shinkai, N., Matsuura, K., Iio, E., Miyaki, T., & Joh, T. (2011). Factors influencing distant recurrence of hepatocellular carcinoma following combined radiofrequency ablation and transarterial chemoembolization therapy in patients with hepatitis C. Cancer Management and Research, 3, 267–272. https://doi.org/10.2147/CMR.S22073

Yap, C. Y., & Aw, T. C. (2010). Liver Function Tests (LFTs). Proceedings of Singapore Healthcare, 19(1), 80–82. https://doi.org/10.1177/201010581001900113

Yarasuri, V. K., Indukuri, G. K., & Nair, A. K. (2019). Prediction of Hepatitis Disease Using Machine Learning Technique. 2019 Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 265–269. https://doi.org/10.1109/I-SMAC47947.2019.9032585

Yue, C., Zhang, C., Ying, C., & Jiang, H. (2022). Reduced serum cholinesterase is an independent risk factor for all-cause mortality in the pediatric intensive care unit. Frontiers in Nutrition, 9. https://doi.org/10.3389/fnut.2022.809449

Recibido: *19 de diciembre de 2025*

Aceptado: *10 de marzo de 2026*

Segovia, José T.: *Ingeniero Aeronáutico y Licenciado en Ciencias y Artes Militares, con estudios de Máster en Bioinformática y Bioestadística, y en Ingeniería de Software e Inteligencia Artificial. Posee más de 20 años de experiencia en gerencia de operaciones, ingeniería y proyectos. Actualmente se desempeña como Gerente General de Certificación de Productos Aeronáuticos (EANSA) y es especialista en el uso de metodologías ágiles e inteligencia artificial.*
https://orcid.org/0009-0004-3187-2335