Machine Learning Vs Deep Learning en la detección del cáncer de hígado

1. Introducción

1.1. Contexto del cáncer de hígado

El cáncer de hígado, particularmente el carcinoma hepatocelular (CHC), es el tipo más común de cáncer hepático, y representa un importante problema de salud pública, debido a su alta mortalidad y su diagnóstico tardío. Según la Organización Mundial de la Salud (OMS), es el tercer cáncer más letal en el mundo, con una tasa de supervivencia a 5 años inferior al 20% en la mayoría de los casos, ya que, a pesar de los avances médicos, la mayoría de los diagnósticos, se realizan en etapas avanzadas, cuando las opciones terapéuticas son limitadas.

El desarrollo del cáncer de hígado, está fuertemente asociado con enfermedades hepáticas crónicas y estilos de vida poco saludables, destacando entre los factores de riesgo más importantes:

Hepatitis B y C crónica: La infección crónica por los virus Hepatitis B (VHB) y Hepatitis C (VHC), es el factor de riesgo más importante para el desarrollo del hepatocarcinoma, llegando al 80% de los casos según la OMS. Estos virus, causan inflamación crónica del hígado, lo que provoca fibrosis y, en muchos casos, cirrosis hepática.

Consumo excesivo de alcohol: El alcoholismo crónico, contribuye significativamente al desarrollo del cáncer de hígado, principalmente a través de la enfermedad hepática alcohólica, ya que, el metabolismo del alcohol en el hígado genera estrés oxidativo, provocando daño celular, inflamación y fibrosis hepática. Además, pacientes con VHB o VHC que consumen alcohol, tienen un riesgo 10 veces mayor de desarrollar CHC, en comparación con personas sin estos factores. Un consumo superior a 40 g/día en hombres y 20 g/día en mujeres, aumenta significativamente el riesgo de cirrosis y cáncer hepático.

Obesidad y diabetes Mellitus: El hígado graso no alcohólico, (NAFLD, por sus siglas en inglés), y su forma avanzada, la esteatohepatitis no alcohólica (NASH), se han convertido en una de las principales causas emergentes de cáncer de hígado, especialmente en países desarrollados. La obesidad y la diabetes generan resistencia a la insulina, lo que promueve la acumulación de grasa en el hígado y el desarrollo de inflamación crónica.

Evolución: NAFLD → NASH → Fibrosis → Cirrosis → Cáncer de hígado.

Actualmente, el 25-30% de la población mundial tiene hígado graso, y de estos, un 20% desarrolla fibrosis avanzada, aumentando el riesgo de CHC.

Impacto del tabaquismo en el desarrollo del cáncer de hígado: El tabaquismo, es un factor de riesgo bien documentado en el desarrollo de múltiples tipos de cáncer. Aunque su impacto es menos directo que el de la hepatitis viral, el alcoholismo o la obesidad, existen estudios que han demostrado una asociación significativa entre el consumo de tabaco y el riesgo de CHC, debido a que, el tabaco, contribuye al desarrollo de inflamación crónica en el hígado, aumentando el riesgo de fibrosis y cirrosis, condiciones que predisponen al CHC.

1.2. La importancia de la detección temprana

El diagnóstico temprano es uno de los mayores desafíos en oncología, ya que en sus primeras fases es asintomático o presenta síntomas no específicos como fatiga, pérdida de peso y malestar abdominal. El problema central, es que, más del 80% de los casos se diagnostican en estadios avanzados, cuando las opciones de tratamiento son limitadas y la mortalidad es alta. Las estrategias actuales de detección incluyen:

Método	Descripción	Ventajas	Limitaciones
Biomarcadores en sangre (Ej: AFP, DCP)	Detección de proteínas asociadas al cáncer en el suero	Prueba sencilla y no invasiva	Baja sensibilidad en etapas tempranas

Método	Descripción	Ventajas	Limitaciones
Imágenes médicas (Ecografía, TAC, RMN)	Visualización de tumores hepáticos mediante radiología	Alta especificidad en tumores avanzados	Alto costo y no es accesible en todas las regiones
Biopsia hepática	Obtención de tejido hepático para análisis histológico	Diagnóstico definitivo	Invasivo, costoso y con riesgo de complicaciones

Principales desventajas:

- **Costo elevado:** Las pruebas de imagen y biopsias hepáticas son costosas y requieren infraestructura especializada.
- **Disponibilidad limitada:** No todos los hospitales tienen acceso a resonancias magnéticas o equipos de biopsia hepática guiada por imagen.
- **Sesgo en la interpretación:** La evaluación de imágenes médicas depende de la experiencia del radiólogo.

Conclusión, no existe un método de detección temprana altamente efectivo y accesible. Por ello, el desarrollo de herramientas basadas en Inteligencia Artificial (IA) representa una alternativa innovadora y prometedora para mejorar el diagnóstico del cáncer de hígado.

1.3. Justificación del estudio: Uso de IA para mejorar la predicción

Debido a las limitaciones indicadas anteriormente, los métodos tradicionales de detección no son suficientes para garantizar una detección temprana. Estos desafíos, han llevado a una necesidad urgente de nuevas herramientas diagnósticas que sean precisas, accesibles y automatizadas. Es aquí, donde entra en juego, la IA, que ha demostrado ser capaz de detectar patrones ocultos en los datos médicos y mejorar la precisión del diagnóstico.

La IA, especialmente los modelos de Machine Learning y Deep Learning, automatiza el análisis de grandes volúmenes de datos médicos y permite, **identificar correlaciones ocultas entre variables clínicas,** que pueden ser ignoradas en análisis convencionales, **optimizar la precisión del diagnóstico,** reduciendo falsos positivos, (pacientes sanos diagnosticados erróneamente con cáncer) y falsos negativos, (casos de cáncer no detectados a tiempo), y **desarrollar herramientas predictivas basadas en tablas,** lo que hace que los modelos sean más accesibles y aplicables en hospitales con menos recursos. Esto lleva, al desarrollo de modelos predictivos accesibles y eficientes, que pueden complementar los métodos tradicionales y mejorar la detección temprana del cáncer de hígado.

1.4. Explicación del PLCO dataset y su origen

El PLCO, (Prostate, Lung, Colorectal, and Ovarian) Cancer Screening Trial Dataset, es una base de datos de acceso restringido, desarrollada por los National Institutes of Health (NIH) de EE.UU. Este conjunto de datos contiene información detallada de cánceres de próstata, pulmón, colorrectal y ovario, pero también incluye datos sobre otros tipos de cáncer, como el de hígado.

Este dataset, se recopiló como parte de un estudio poblacional diseñado, para evaluar la eficacia de diferentes estrategias de detección temprana de cáncer. El dataset, no se basa en registros hospitalarios tradicionales, sino que se construyó, a partir de formularios voluntarios completados por miles de pacientes en un seguimiento a largo plazo.

Los participantes, rellenaban encuestas médicas periódicas que incluían; Información demográfica, (edad, género, etnia), estilos de vida, (tabaquismo, consumo de alcohol, actividad física), antecedentes médicos, (historial de cáncer, enfermedades hepáticas, cirrosis), resultados de análisis clínicos y biomarcadores. El objetivo del estudio, era monitorear a los participantes, para identificar quiénes desarrollaban cáncer y quiénes no.

Este dataset, es único en comparación con otras bases de datos médicas tradicionales porque:

Es un estudio poblacional de seguimiento, que no se limita a datos de hospitales, sino que rastrea la salud de los pacientes a lo largo del tiempo. Esto permite, estudiar la evolución de factores de riesgo antes del desarrollo del cáncer.

Incluye, datos detallados sobre múltiples cánceres, incluyendo el de hígado, haciendo posible la posibilidad de analizar patrones comunes, en pacientes con diferentes tipos de cáncer, y permite evaluar, cómo los factores de riesgo evolucionan a lo largo del tiempo.

Es un dataset de alta calidad y validado. Los datos provienen de una institución médica de prestigio (NIH), que se han utilizado en múltiples investigaciones científicas publicadas en revistas médicas.

Aun así, el dataset, presenta numerosos desafíos, como por ejemplo, la existencia de **valores faltantes,** debido a respuestas voluntarias incompletas, y, también, puede presentar **desbalanceamiento** de clases, (más participantes sanos que enfermos).

1.5. Definición de la variable liver_cancer como target principal

El objetivo central de esta investigación, es predecir la presencia de cáncer de hígado, y para ello, se define la variable '***liver_cancer',*** como el target principal en los modelos de Machine Learning y Deep Learning.

Definición de la variable:

- **liver_cancer = 1** → Paciente ha sido diagnosticado con cáncer de hígado.
- **liver_cancer = 0** → Paciente sin diagnóstico de cáncer de hígado.

Esta variable, permite formular el problema como una tarea de clasificación binaria, en la que los modelos de IA, deben aprender a distinguir entre pacientes con y sin cáncer hepático.

2. Metodología general del estudio

En este contexto, este estudio busca comparar dos enfoques distintos dentro de la Inteligencia Artificial para la detección del cáncer de hígado: Machine Learning y Deep Learning.

El enfoque de Machine Learning, se basa en algoritmos supervisados que aprenden de patrones en los datos y generan predicciones a partir de relaciones identificadas entre las variables clínicas. En contraste, el enfoque de Deep Learning, utiliza redes neuronales artificiales, para analizar los datos de una manera más compleja, siendo capaz de descubrir correlaciones profundas que pueden ser difíciles de captar con modelos tradicionales.

Ambos enfoques, presentan características únicas, ventajas y limitaciones que los hacen adecuados en distintos contextos. Mientras que Machine Learning es más fácil de interpretar y requiere menos datos, Deep Learning es capaz de capturar patrones más complejos, aunque a expensas de una mayor demanda computacional y una menor facilidad de interpretación. La meta de este estudio, es **determinar cuál de los dos enfoques ofrece mejores resultados en la predicción del cáncer de hígado** a partir de datos clínicos, evaluando su rendimiento a través de métricas estandarizadas.

2.1. Explicación de los dos bloques del estudio

2.1.1. Machine Learning

Los algoritmos de Machine Learning, han sido utilizados en la medicina para la clasificación y predicción de enfermedades. En este estudio, se emplean modelos de aprendizaje supervisado, para analizar la presencia de cáncer de hígado, a partir de datos clínicos obtenidos en el PLCO Dataset.

El principio fundamental de Machine Learning, es el uso de modelos estadísticos y matemáticos que pueden identificar patrones en los datos sin necesidad de una programación explícita de reglas. Estos modelos, se entrenan con datos históricos en los que se conoce el diagnóstico de los pacientes y, posteriormente, pueden predecir la probabilidad de que un nuevo paciente desarrolle cáncer de hígado.

Para este estudio, se han seleccionado varios algoritmos de Machine Learning ampliamente utilizados en la clasificación médica: regresión logística, Random Forest, Support Vector Machines (SVM) y Gradient Boosting (XGBoost). Cada uno de estos algoritmos tiene sus propias características y es evaluado en función de su precisión, capacidad de generalización y facilidad de interpretación.

La regresión logística, es el modelo más simple y es utilizado como línea base en este estudio, ya que proporciona una forma rápida de evaluar la relación entre las variables clínicas y la variable objetivo '***liver_cancer'***. Por otro lado, Random Forest, es un modelo basado en árboles de decisión que permite capturar relaciones no lineales y reducir el riesgo de sobreajuste, mediante la combinación de múltiples árboles. SVM, es un modelo que busca encontrar un hiperplano óptimo, para separar las clases de pacientes con y sin cáncer, siendo particularmente útil cuando los datos tienen una estructura compleja. Finalmente, los modelos basados en boosting, como XGBoost, utilizan un enfoque iterativo, para mejorar progresivamente la capacidad predictiva del modelo, permitiendo una mejor precisión en comparación con otros algoritmos tradicionales.

Uno de los aspectos clave, del uso de Machine Learning en este estudio, es la necesidad de realizar un preprocesamiento adecuado de los datos. Aplicaré técnicas como normalización de variables, eliminación de valores atípicos y balanceo de clases, con el fin de garantizar que los modelos no se vean afectados por sesgos en la distribución de los datos. Además, realizaré una selección de características para descartar variables redundantes y mejorar la eficiencia de los algoritmos.

A pesar de sus ventajas, el uso de Machine Learning en la detección del cáncer de hígado también presenta desafíos. La necesidad de definir manualmente las características más relevantes para la clasificación, puede ser una limitación en comparación con el Deep Learning, que es capaz de aprender estas representaciones de manera automática. Además, algunos modelos de Machine Learning pueden no capturar relaciones altamente complejas en los datos, lo que puede afectar su capacidad de generalización.

2.1.2. Deep Learning, redes neuronales

El Deep Learning, utiliza redes neuronales artificiales para analizar datos de una manera más avanzada. A diferencia de los algoritmos de Machine Learning, las redes neuronales pueden aprender automáticamente representaciones complejas de los datos sin necesidad de definir manualmente las características a considerar en el modelo.

En este estudio, se ha implementado una Red Neuronal Artificial (ANN), para la predicción del cáncer de hígado, utilizando la biblioteca TensorFlow/Keras. La arquitectura de la red neuronal está compuesta por varias capas densas interconectadas, donde cada capa transforma los datos de entrada en representaciones más abstractas antes de producir una salida final.

Una de las principales ventajas del Deep Learning, es su capacidad para manejar datos complejos y detectar patrones ocultos que pueden no ser evidentes para los modelos tradicionales de Machine Learning. Sin embargo, esta mayor capacidad predictiva también conlleva mayores requerimientos computacionales y una menor facilidad de interpretación, lo que puede ser un obstáculo en aplicaciones médicas, donde la transparencia del modelo es crucial.

Para mejorar el rendimiento de la red neuronal, en este estudio se han aplicado técnicas avanzadas como Batch Normalization y Dropout, que permiten estabilizar el entrenamiento y reducir el riesgo de sobreajuste. Además, se ha utilizado Keras Tuner para optimizar la arquitectura del modelo, ajustando parámetros como el número de neuronas, la tasa de aprendizaje y la cantidad de capas ocultas.

A pesar de sus beneficios, el uso de Deep Learning en este estudio también presenta limitaciones. Dado que los modelos de redes neuronales requieren grandes volúmenes de datos para su entrenamiento, su rendimiento puede verse afectado si los datos disponibles no son suficientes. Además, la falta de interpretación en las decisiones del modelo puede hacer que su aplicación en entornos clínicos sea más complicada, en comparación con los modelos de Machine Learning tradicionales.

2.2. Estrategia de comparación entre ambos enfoques

Dado que ambas metodologías tienen sus propias fortalezas y debilidades, este estudio, busca comparar ambos enfoques de manera rigurosa, para determinar cuál ofrece un mejor rendimiento en la predicción del cáncer de hígado.

Con este objetivo, se ha diseñado una estrategia basada en métricas de evaluación estandarizadas en clasificación binaria. Entre las métricas utilizadas se encuentran precisión, sensibilidad, especificidad y ROC-AUC Score, cada una proporcionando información clave sobre la capacidad predictiva de los modelos.

El análisis de resultados, no solo se centrará en la precisión del modelo, sino también en su capacidad para detectar correctamente los casos positivos de cáncer de hígado, sensibilidad, lo cual es crucial en aplicaciones médicas. Un modelo con una precisión alta pero una sensibilidad baja, podría no ser útil en la práctica, ya que podría fallar en identificar pacientes enfermos.

Además, se considerará la interpretabilidad del modelo, ya que en aplicaciones médicas, es fundamental, que los especialistas puedan entender cómo el modelo llega a sus decisiones. En este sentido, los modelos de Machine Learning, como Random Forest, permiten obtener la importancia de cada variable en la predicción, mientras que en Deep Learning se explorarán métodos como SHAP (Shapley Additive Explanations), para mejorar la explicabilidad del modelo.

Con esta estrategia de comparación, se busca no solo determinar cuál modelo tiene mejor rendimiento, sino también identificar cuál sería más factible para su implementación en la práctica médica.

3. Técnicas de preprocesamiento aplicadas

El preprocesamiento de datos, es una fase crítica en cualquier proyecto de Machine Learning o Deep Learning, especialmente cuando se trabaja con datos clínicos, como los del PLCO Dataset. En este estudio, el preprocesamiento, no solo se enfoca en limpiar y transformar los datos, sino también en optimizar la calidad de la información para garantizar que los modelos puedan aprender de manera eficiente y precisa.

Se han implementado, diversas técnicas para mejorar la representatividad y la utilidad del dataset, incluyendo Feature Engineering, imputación de valores faltantes, balanceo de clases y la creación de múltiples versiones del dataset para evaluar el impacto de diferentes transformaciones.

Sin embargo, antes de preprocesar nuestro dataset, dividí mi dataset en conjuntos de entrenamiento y prueba debido a las siguientes razones:

- Evitar **data leakage o fuga de información**, que ocurre cuando la información del conjunto de prueba se utiliza indirectamente durante el entrenamiento. Esto puede llevar, a una evaluación optimista e irreal del rendimiento del modelo. Al dividir los datos al principio, aseguramos que el conjunto de prueba permanezca completamente independiente del proceso de entrenamiento.
- Evaluar el **rendimiento real** del modelo. La división temprana entre entrenamiento y prueba, asegura que el modelo solo vea los datos de entrenamiento, durante el proceso de ajuste. El conjunto de prueba, debe usarse solo, para la evaluación final del modelo, de modo que refleje el rendimiento del modelo en datos “nuevos” e invisibles para el modelo durante el entrenamiento.
- **Prevenir sesgos** en el modelo. Si se realiza el preprocesamiento y balanceo antes de la división de los datos, los parámetros calculados en los datos de entrenamiento, (medias, transformaciones,…), podrían verse afectados por los datos de prueba. Esto sesgaría el modelo, ya que tendría acceso a información sobre el conjunto de prueba.

3.1. Feature Engineering y selección de variables

En este estudio, este proceso se centro en seleccionar y descartar variables con el objetivo de mejorar la capacidad predictiva del modelo para la detección del cáncer de hígado. Para ello, se realizó un análisis detallado de cada sección del

dataset proporcionado por la NIH, clasificando las variables según su relevancia, eliminando información redundante o irrelevante y generando nuevas representaciones de datos cuando fue necesario.

En primer lugar, se exploró el dataset original, identificando el número de observaciones y variables disponibles. Se realizó una clasificación de las variables en diferentes secciones basadas en la documentación del PLCO dataset, permitiendo un análisis sistemático de cada grupo de características. Se examinó la distribución de los valores, la cantidad de datos faltantes y la correlación entre variables, asegurando que solo aquellas características con valor informativo significativo fueran consideradas en el modelo final. Además, se eliminaron variables con valores únicos o con altos porcentajes de valores nulos, ya que estas no aportaban información relevante para la predicción del cáncer de hígado.

A lo largo del análisis, se tomó la decisión de descartar variables redundantes o que no proporcionaban información adicional significativa. Por ejemplo, en la sección de datos demográficos, se encontró una alta correlación entre las variables "age" y "agelevel", ambas representando la edad de los participantes. Se optó por conservar "agelevel", ya que representaba la edad de manera categórica y contenía menos valores distintos, facilitando su tratamiento en modelos de Machine Learning. De manera similar, en la sección de características del cáncer, se eliminaron variables como "liver_behavior", "liver_grade" y "liver_morphology", ya que este estudio se enfoca en la predicción de la aparición de cáncer de hígado, y no en la clasificación de la enfermedad una vez diagnosticada.

Otro aspecto clave del proceso fue la eliminación de datos no relevantes para la predicción, como los identificadores (plco_id), variables administrativas (bq_returned, bq_compdays) y variables de consentimiento (reconsent_outcome). También se eliminaron registros con información incompleta, como aquellos donde el formulario de salud fue rechazado debido a un diagnóstico previo de cáncer, garantizando que el análisis estuviera basado en datos representativos de la población en riesgo.

Como resultado del proceso de feature engineering y selección de variables, se construyó una versión optimizada del dataset, que contenía solo las variables relevantes para la predicción de liver_cancer. Este preprocesamiento permitió reducir la dimensionalidad del dataset y mejorar la eficiencia de los modelos de Machine Learning y Deep Learning, asegurando que los datos utilizados en el entrenamiento sean de alta calidad y libres de redundancias o información irrelevante.

3.2. Imputación de valores faltantes

La presencia de valores faltantes en datasets clínicos, es un desafío común, que puede afectar significativamente la calidad de los modelos predictivos. En el caso del PLCO dataset, los datos fueron recolectados a lo largo de varios años mediante formularios voluntarios, lo que incrementa la posibilidad, de que existan variables incompletas, debido a respuestas omitidas o errores en la recopilación.

Dado que el objetivo de este estudio, es la predicción del cáncer de hígado a partir de la variable *'liver_cancer'*, es fundamental, que los datos estén lo más completos posible para mejorar la precisión de los modelos de Machine Learning y Deep Learning. Para ello, se han aplicado distintas técnicas de imputación de valores faltantes, las cuales pueden clasificarse en imputación simple e imputación múltiple.

Cada uno de estos métodos se estudió, creando y entrenando diversos modelos, para analizar y seleccionar la técnica que arrojó mejores resultados. Para la toma de esa decisión, se crearon matrices de confusión y se analizaron métricas como el error cuadrático, R^2, y la raíz de la desviación cuadrática media, RMSE.

3.2.1. Imputación simple

Es una de las estrategias más utilizadas para tratar valores faltantes, que consiste en reemplazar los valores ausentes, con un ***único*** estimador basado en la distribución de los datos observados. Este método, es adecuado, cuando la cantidad de valores faltantes es baja y cuando la variable afectada, no tiene una fuerte dependencia con otras variables. Aunque es una solución rápida y sencilla, tiene la limitación de no capturar la incertidumbre en la imputación ni la relación entre variables.

3.2.2. Imputación múltiple

Se han explorado, diversas técnicas de imputación simple. Sin embargo, cada uno de estos métodos, ha demostrado ser útil en distintos escenarios, pero comparten una limitación clave. Todas estas técnicas, generan un único valor de imputación para cada observación faltante, sin considerar la incertidumbre de la predicción.

Para abordar esta limitación, se aplicó imputación múltiple mediante ecuaciones encadenadas (MICE, Multiple Imputation by Chained Equations). Este método, permite, estimar ***múltiples valores*** posibles para cada celda faltante, lo que ayuda a reflejar mejor la variabilidad y la estructura subyacente de los datos.

3.2.3. Procedimiento en el estudio

Antes de aplicar cualquier método de imputación, se realizó un análisis exploratorio para determinar el porcentaje de valores faltantes en cada variable. Este análisis permitió clasificar las variables en función del porcentaje de datos ausentes y seleccionar la técnica de imputación más adecuada.

Los resultados mostraron que algunas variables contenían un porcentaje elevado de valores faltantes, mientras que otras presentaban ausencias en menos del 1% de las observaciones. A partir de estos resultados, se implementaron diferentes estrategias de imputación.

Tras aplicar diferentes métodos de imputación a los valores faltantes en el PLCO dataset, se realizó una evaluación detallada del impacto de cada técnica en la estructura de los datos y en el rendimiento de los modelos de Machine Learning. Para ello, se analizaron los cambios en la correlación entre variables tras la imputación y se compararon los resultados en términos de error cuadrático medio (RMSE) y coeficiente de determinación (R^2) en varios modelos de predicción.

El primer aspecto que se evaluó, fue cómo la imputación afectó la relación entre las variables del dataset. Se observó, que la imputación basada en MICE, generó la mayor variación en la matriz de correlación, lo que sugiere, que este método, reestructuró en mayor medida, la relación entre las variables en comparación con las demás técnicas. Por otro lado, la imputación mediante KNN y regresión lineal, produjo un menor grado de alteración en la estructura de correlaciones, mientras que la imputación por media y mediana mantuvo las correlaciones más cercanas al dataset original. Este hallazgo es relevante, ya que cambios significativos en la estructura de correlaciones pueden afectar la capacidad de los modelos para aprender patrones relevantes en los datos.

En la evaluación del rendimiento de los modelos predictivos, se encontró que las diferencias en RMSE y R^2 entre los métodos de imputación fueron mínimas en modelos lineales como regresión lineal y KNN regression, lo que indica que ninguno de los métodos tuvo un impacto drástico en la precisión de la predicción de liver_cancer. Sin embargo, en modelos más avanzados como Random Forest y XGBoost, se observó que el rendimiento fue idéntico independientemente del método de imputación empleado, con valores de RMSE de 0.0058 y R^2 de 0.9743 en todas las imputaciones. Este resultado sugiere que, en modelos más complejos y robustos, la imputación de valores faltantes no tuvo un impacto significativo en la capacidad del modelo para predecir la variable objetivo.

Otro aspecto importante a considerar fue el costo computacional de cada técnica. Mientras que MICE y KNN, requirieron un tiempo considerable para la imputación, debido a su naturaleza iterativa y dependiente de múltiples cálculos, la imputación con media y mediana fue la más eficiente en términos de procesamiento. Esta diferencia, en tiempos de cómputo, se vuelve relevante cuando se trabaja con datasets grandes, donde métodos más avanzados pueden volverse computacionalmente costosos sin aportar mejoras sustanciales en el rendimiento del modelo.

Finalmente, el análisis demostró que la imputación mediante media y mediana preservó mejor la estructura de los datos sin generar grandes variaciones en la distribución original de las variables. Aunque MICE se basa en un enfoque más sofisticado y estadísticamente sólido, su aplicación en este caso particular no ofreció ventajas significativas en términos de precisión

predictiva. Además, al generar mayores cambios en la correlación entre variables, introdujo una mayor incertidumbre en la estructura del dataset, lo que podría afectar la hora de interpretar los modelos en un contexto clínico.

3.2.4. Conclusión

Tras evaluar los diferentes métodos de imputación en términos de estabilidad de los datos, impacto en las correlaciones, rendimiento en modelos de Machine Learning y eficiencia computacional, se concluyó que la imputación con media y mediana era la opción más adecuada para este estudio.

Una de las principales razones que justificaron esta elección fue su simplicidad y eficiencia, lo que permitió una imputación rápida y efectiva sin alterar significativamente la estructura original de los datos. Este aspecto es especialmente relevante en estudios de Machine Learning en medicina, donde los tiempos de cómputo pueden ser un factor limitante al trabajar con grandes volúmenes de datos clínicos.

Otro factor importante fue el análisis del RMSE y el R^2 en modelos de Machine Learning. Se encontró que las diferencias en rendimiento entre los métodos de imputación eran prácticamente insignificantes, especialmente en modelos avanzados como Random Forest y XGBoost, donde todos los métodos generaron resultados idénticos en términos de precisión. Esto sugiere que utilizar una técnica más compleja como MICE no ofrecía ventajas significativas en este caso y que métodos más simples podían lograr el mismo nivel de precisión sin la complejidad adicional.

Finalmente, desde una perspectiva de aplicabilidad clínica, la imputación con media y mediana ofrece un enfoque más estable y replicable, lo que facilita su implementación en futuros estudios o sistemas de predicción. En entornos médicos, es fundamental que las técnicas utilizadas para el tratamiento de datos sean transparentes y fáciles de interpretar, ya que los resultados del modelo pueden influir en decisiones clínicas críticas.

En conclusión, se decidió utilizar la imputación con media y mediana como método final, ya que demostró ser una técnica eficiente, estable y confiable que preserva mejor la estructura de los datos sin introducir cambios significativos en las correlaciones entre variables. La decisión final se basó en un equilibrio entre precisión, estabilidad y eficiencia computacional, priorizando una solución que optimice la calidad de los datos sin comprometer la robustez del modelo predictivo.

3.4. Data transformation

La transformación de datos es una fase esencial en la preparación del conjunto de datos para el modelado de Machine Learning. Su propósito es mejorar la calidad de los datos, optimizar la capacidad predictiva del modelo y garantizar la coherencia entre los conjuntos de entrenamiento y prueba. En este estudio, se implementaron diversas estrategias para seleccionar y transformar características, asegurando que los modelos predictivos fueran más precisos, estables y eficientes.

Uno de los primeros pasos consistió en la selección de características, donde se identificaron las variables más relevantes mediante técnicas como Mutual Information, Pearson y Spearman correlation. Estas métricas permitieron evaluar la relación de cada variable con la variable objetivo, eliminando aquellas con baja importancia o alta redundancia. Al reducir el número de características irrelevantes, se consiguió mejorar la eficiencia computacional del modelo sin comprometer su rendimiento, evitando el sobreajuste y asegurando que el modelo se enfocara en las variables con mayor impacto en la predicción del cáncer de hígado.

Además de la selección de variables, se aplicaron diversas transformaciones matemáticas para mejorar la distribución de los datos y hacerlos más adecuados para los algoritmos de Machine Learning. Algunas variables fueron sometidas a transformaciones logarítmicas para reducir la asimetría en su distribución, mientras que otras fueron elevadas al cuadrado o a la cuarta potencia para resaltar patrones no lineales. En otros casos, la raíz cuadrada y la raíz cuarta fueron utilizadas para suavizar la variabilidad de ciertas características. Estas transformaciones ayudaron a que los modelos captaran de manera más efectiva las relaciones complejas en los datos, mejorando la estabilidad del entrenamiento y reduciendo la influencia de valores extremos.

Para evaluar la efectividad de estas transformaciones, se compararon dos enfoques distintos: uno en el que se eliminaron las características con menor importancia y otro en el que se aplicaron transformaciones a todas las variables relevantes. Ambos conjuntos de datos fueron utilizados para entrenar modelos de Random Forest, analizando métricas clave como precisión, F1-score, recall y AUC-ROC. Los resultados indicaron que ambas estrategias tenían un rendimiento similar en términos de predicción, pero el modelo que utilizaba un menor número de variables era más eficiente computacionalmente. Esto sugiere que una transformación adecuada de los datos permite obtener modelos más ligeros sin perder precisión en las predicciones.

Un aspecto fundamental del proceso de transformación fue garantizar la coherencia entre el conjunto de entrenamiento y el conjunto de prueba. Para evitar sesgos y asegurar que los modelos fueran evaluados en condiciones realistas, se aplicaron las mismas transformaciones a los datos de prueba. Esta estrategia evitó problemas como data leakage, donde el modelo puede beneficiarse inadvertidamente de información que no estaría disponible en un escenario real. Gracias a esta consistencia en el preprocesamiento, los resultados obtenidos en la fase de prueba reflejan con mayor fidelidad la capacidad de generalización del modelo.

Los resultados de esta etapa demostraron que la transformación de datos no solo optimizó el rendimiento del modelo, sino que también facilitó su implementación al reducir la complejidad del conjunto de datos. La eliminación de variables irrelevantes permitió construir modelos más eficientes, mientras que la aplicación de transformaciones matemáticas mejoró la estabilidad del entrenamiento y la capacidad de los modelos para identificar patrones en los datos. En última instancia, este enfoque permitió desarrollar un sistema predictivo más robusto, con mayor capacidad para detectar casos de cáncer de hígado y mejorar su aplicabilidad en entornos médicos.

3.5. Balanceo de datos (SMOTE, Undersampling, Oversampling)

Uno de los desafíos más importantes en el desarrollo de modelos de Machine Learning para la predicción del cáncer de hígado es el desbalanceamiento en el dataset. En problemas médicos, es común que los datos de pacientes con la enfermedad sean significativamente menores en comparación con los pacientes sanos. Este desequilibrio puede generar modelos que favorezcan la clasificación de la clase mayoritaria (no cáncer) y disminuyan la capacidad de detección de casos positivos (cáncer).

Para abordar este problema, en este estudio se implementaron tres técnicas principales de balanceo de datos. Cada método fue aplicado y evaluado mediante la creación de modelos de Machine Learning, con el objetivo de determinar que técnica mejoraba la capacidad del modelo para detectar el cáncer de hígado, sin introducir sesgos en los datos.

- **SMOTE (Synthetic Minority Over-sampling Technique):** técnica de oversampling que genera nuevas instancias sintéticas de la clase minoritaria en lugar de simplemente duplicar observaciones existentes. Utiliza la interpolación de características entre ejemplos reales de la clase minoritaria, lo que ayuda a mejorar la capacidad del modelo para generalizar.
- **Oversampling:** Duplica aleatoriamente observaciones de la clase minoritaria para aumentar su representación.
- **Undersampling (NearMiss):** Reduce la cantidad de ejemplos de la clase mayoritaria para alcanzar un equilibrio en la distribución de clases.

3.5.1. Procedimiento en el estudio

Para abordar este problema, se implementaron técnicas de balanceo de datos utilizando herramientas de la biblioteca imblearn (imbalanced-learn). Se aplicaron las técnicas anteriormente mencionadas, con el objetivo de determinar cuál de estos métodos lograba mejorar la capacidad predictiva del modelo sin introducir sobreajuste ni pérdida de información valiosa. Se diseñó una metodología basada en el uso de validación cruzada estratificada, garantizando que cada modelo fuera evaluado en diferentes particiones del dataset y no en un único conjunto de entrenamiento y prueba, lo que permitió obtener métricas más confiables y generalizables.

El procedimiento de balanceo se desarrolló en varios pasos. Primero, se separaron las características del dataset (X) y la variable objetivo (y). Luego, se aplicaron las diferentes técnicas de balanceo al dataset de entrenamiento, lo que permitió

generar un nuevo conjunto de datos con una distribución de clases equitativa. Para asegurar, que los cambios introducidos por el balanceo, no afectaran la estructura de los datos, se realizaron comparaciones visuales y estadísticas entre el dataset original y los datasets balanceados. Esto incluyó análisis de correlaciones, histogramas de distribución y comparación de métricas en modelos entrenados con cada método de balanceo.

El código implementado permitió automatizar la comparación entre los métodos de balanceo y seleccionar el más adecuado en función de métricas clave como accuracy, balanced accuracy, precision, recall y F1-score. Se utilizó un modelo base de Random Forest como clasificador estándar para evaluar cada técnica. La función de evaluación desarrollada en el código no solo permitió obtener los resultados de cada modelo, sino que también incluyó la generación de matrices de confusión, lo que facilitó la interpretación visual del desempeño de cada método en términos de clasificación correcta e incorrecta de las clases.

En resumen, el procedimiento aplicado en este estudio se enfocó en: identificar el problema de desbalanceo, aplicar y evaluar diferentes estrategias de corrección y comparar los modelos resultantes para seleccionar la mejor técnica. Gracias a este enfoque estructurado, se logró mejorar la capacidad del modelo para detectar casos de cáncer de hígado, asegurando que los resultados obtenidos sean confiables y aplicables en la predicción de enfermedades mediante Machine Learning y Deep Learning.

3.5.2. Conclusión

Tras la aplicación de SMOTE, oversampling y undersampling, así como la evaluación de su impacto en el rendimiento de los modelos de Machine Learning, se obtuvieron resultados altamente positivos en términos de accuracy, balanced accuracy, precision, recall y F1-score.

El análisis de los resultados obtenidos en la evaluación de los modelos balanceados reveló que SMOTE fue la técnica más efectiva para corregir el desbalanceamiento de clases en el dataset y mejorar la capacidad predictiva del modelo en la detección de pacientes con cáncer de hígado. Aunque oversampling y undersampling también lograron mejorar el rendimiento del modelo, SMOTE destacó principalmente por su impacto en la sensibilidad, que es una de las métricas más importantes en problemas médicos.

Los resultados mostraron que, tras aplicar SMOTE, el recall alcanzó el 100%, lo que indica que el modelo pudo detectar absolutamente todos los casos positivos sin errores. Esto es fundamental, ya que antes del balanceo, el modelo mostraba un sesgo hacia la clase mayoritaria (liver_cancer = 0), resultando en una detección ineficaz de los pacientes con cáncer.

El aumento en la sensibilidad se debe a la manera en que SMOTE genera ejemplos sintéticos. En lugar de simplemente duplicar observaciones existentes, SMOTE crea nuevos datos sintéticos interpolando entre puntos reales de la clase minoritaria. Esto permite que el modelo aprenda patrones más representativos de los casos positivos sin caer en problemas de sobreajuste.

En comparación con undersampling, SMOTE no eliminó datos valiosos. La técnica de undersampling redujo el tamaño del dataset, eliminando instancias de la clase mayoritaria, lo que puede llevar a una pérdida de información crítica. Aunque undersampling logró también resultados perfectos en este estudio, su impacto en datasets más grandes, podría ser negativo debido a la reducción de datos de entrenamiento.

En conclusión, SMOTE fue la mejor técnica de balanceo para este estudio, ya que permitió crear un modelo predictivo más robusto, con una mejor capacidad para detectar casos positivos de liver_cancer, sin comprometer la estabilidad y generalización del modelo, preservando la estructura del dataset sin eliminar información valiosa ni generar sobreajuste.

4. Algoritmos de Machine Learning

El desarrollo del modelo, para la predicción del cáncer de hígado, se llevó a cabo mediante la evaluación comparativa de varios algoritmos de Machine Learning, con el objetivo de identificar, cuál tenía un mejor desempeño en términos de precisión, sensibilidad y área bajo la curva ROC. Se utilizaron algoritmos como regresión logística, Random Forest, XGBoost, decision tree,

AdaBoost y SVM, aplicando técnicas de validación cruzada para asegurar que los resultados fueran representativos y generalizables.

4.1. Creación de modelo y comparación de algoritmos

El primer paso fue definir una serie de algoritmos de Machine Learning para evaluar su desempeño en la predicción del cáncer de hígado. Se seleccionaron modelos de distintos enfoques, incluyendo modelos lineales, árboles de decisión y métodos de ensemble learning. Los modelos probados inicialmente fueron:

- **Regresión Logística (LogisticRegression):** Modelo lineal clásico utilizado en clasificación binaria.
- **Árbol de Decisión (DecisionTreeClassifier):** Modelo basado en reglas de decisión.
- **Random Forest (RandomForestClassifier):** Ensamble de múltiples árboles de decisión para mejorar la generalización.
- **SVM (Support Vector Machine) (LinearSVC):** Modelo lineal que encuentra el hiperplano óptimo para separar las clases.
- **AdaBoost (AdaBoostClassifier):** Algoritmo de boosting que mejora la predicción combinando modelos débiles.
- **XGBoost (XGBClassifier):** Algoritmo de boosting basado en árboles de decisión optimizado para grandes volúmenes de datos.

El entrenamiento de los modelos se realizó, utilizando el dataset donde los datos habían sido previamente procesados para eliminar el problema de desbalanceamiento de clases. Se utilizó un esquema de validación cruzada estratificada, asegurando que los datos de entrenamiento y prueba conservaran la misma proporción de clases en cada iteración y se evaluaron varias métricas clave, incluyendo accuracy, recall, precision y F1-score, con énfasis en la sensibilidad, dado que el objetivo del modelo es maximizar la detección de pacientes con cáncer de hígado.

4.1.1. Métricas claves

En problemas de clasificación binaria como la detección del cáncer de hígado, la elección de las métricas adecuadas es fundamental para evaluar la calidad del modelo. No todas las métricas tienen el mismo peso, y su interpretación puede cambiar dependiendo de si el dataset está balanceado o desbalanceado.

Para este estudio, se priorizaron tres métricas clave: Accuracy (precisión global), recall (sensibilidad o tasa de verdaderos positivos), ROC-AUC (área bajo la curva ROC). Cada una de estas métricas ofrece información diferente sobre el desempeño del modelo, y su correcta interpretación permite tomar decisiones sobre qué modelo es más adecuado para la detección del cáncer de hígado.

Accuracy (Precisión Global): Mide la proporción de predicciones correctas en relación con el total de casos. Aunque accuracy es una métrica estándar en problemas de clasificación, puede ser engañosa en datasets desbalanceados. Si la mayoría de los pacientes no tienen cáncer, un modelo puede obtener una accuracy alta simplemente prediciendo siempre la clase mayoritaria (liver_cancer = 0), sin identificar correctamente los casos positivos.

Recall (Sensibilidad o Tasa de Verdaderos Positivos): Es la métrica más importante en este estudio, ya que mide la capacidad del modelo para detectar correctamente los casos positivos de cáncer. En términos médicos, un recall bajo significaría que el modelo está fallando en identificar a pacientes enfermos, lo que puede ser crítico.

ROC-AUC (Área Bajo la Curva ROC): Evalúa la capacidad del modelo para diferenciar entre clases (liver_cancer = 0 y liver_cancer = 1). Un valor alto de AUC cercano a 1 indica que el modelo tiene una alta capacidad para distinguir correctamente entre pacientes sanos y enfermos.

4.2. Resultados iniciales y optimización de hiperparámetros

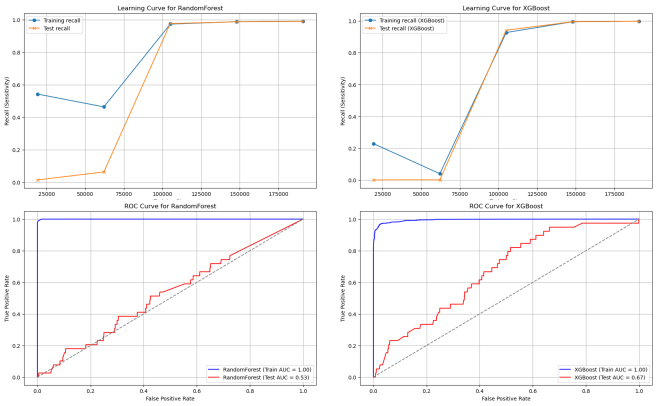
Antes de ajustar los hiperparámetros, se evaluaron los modelos en su configuración estándar para establecer una línea base de comparación. Los resultados iniciales de la evaluación mostraron que decision tree tuvo un desempeño decente, pero mostró mayor varianza en los resultados, lo que sugiere sobreajuste y regresión logística y SVC no lograron capturar la

complejidad del problema, presentando menor recall. Random forest y XGBoost tenían los mejores valores en precisión y recall, por lo que se decidió realizar un ajuste de hiperparámetros para optimizar estos modelos.

Para **Random Forest**, se ajustaron parámetros como: **Número de árboles, profundidad máxima, peso de clases.**

En el caso de **XGBoost**, se optimizaron parámetros como: **learning_rate, n_estimators, scale_pos_weight.**

Tras la aplicación de RandomizedSearchCV y GridSearchCV para afinar los hiperparámetros, se observaron mejoras significativas en las métricas clave, particularmente en recall y AUC-ROC. La optimización permitió que ambos modelos fueran más eficientes en la identificación de casos positivos sin comprometer la precisión general del modelo.



Análisis de Resultados: Random Forest vs XGBoost

En el caso de Random Forest, el recall se incrementó considerablemente tras la optimización, alcanzando valores cercanos al 99% en el conjunto de entrenamiento. Sin embargo, la curva de aprendizaje muestra que el modelo presentó un comportamiento algo inestable con tamaños de muestra pequeños, pero conforme aumentó la cantidad de datos, el recall mejoró significativamente y alcanzó un nivel óptimo.

No obstante, la curva ROC-AUC del modelo optimizado evidenció un sobreajuste considerable, ya que en el conjunto de entrenamiento el AUC alcanzó 1.00 (clasificación perfecta), mientras que en el conjunto de prueba descendió hasta 0.53, lo que sugiere una baja capacidad de generalización del modelo.

Por otro lado, XGBoost también mostró una mejora significativa en recall, logrando un desempeño sobresaliente en el conjunto de entrenamiento. Su curva de aprendizaje demuestra que el modelo aprende de manera progresiva y se estabiliza rápidamente en valores altos de recall tanto para los datos de entrenamiento como de prueba.

La curva ROC-AUC de XGBoost, aunque también evidencia cierto sobreajuste, tuvo un mejor desempeño en el conjunto de prueba en comparación con Random Forest, con un AUC-ROC de 0.67, lo que indica una mayor capacidad de diferenciación entre clases.

4.3. Conclusiones

Random Forest y XGBoost fueron los modelos con mejor desempeño, destacando por su capacidad para detectar correctamente los casos positivos sin sacrificar precisión. Se demostró que la sensibilidad es un factor clave en la evaluación de estos modelos, ya que en problemas médicos es más importante detectar correctamente los casos positivos que minimizar falsos positivos.

El uso de validación cruzada permitió garantizar que los resultados fueran estables y reproducibles, evitando sobreajuste en el modelo seleccionado. XGBoost emergió como el mejor modelo, con una combinación óptima de recall y AUC-ROC, asegurando una mejor detección de pacientes con cáncer de hígado. Sin embargo, random forest sigue siendo una alternativa confiable y más interpretable, aspecto fundamental en entornos médicos, aunque con una ligera desventaja en recall.

5. Algoritmos de Deep Learning

El uso de redes neuronales artificiales, en la predicción del cáncer de hígado, permite capturar patrones complejos y relaciones no lineales en los datos médicos, mejorando la precisión diagnóstica con respecto a los modelos tradicionales de Machine Learning.

En esta sección, exploraremos el desarrollo de un modelo de Deep Learning, detallando el proceso desde la construcción de la arquitectura de la red hasta su evaluación con métricas avanzadas. Mediante este enfoque, buscamos construir un modelo robusto, capaz de mejorar la detección temprana del cáncer de hígado y contribuir al desarrollo de herramientas de apoyo en el diagnóstico médico.

5.1. Creación de redes neuronales artificiales (ANN)

Las redes neuronales artificiales (ANN), son modelos computacionales inspirados en la estructura del cerebro humano, capaces de aprender patrones complejos a partir de los datos. En este estudio, se ha desarrollado, un modelo basado en una ANN, para predecir la aparición del cáncer de hígado a partir de un conjunto de características clínicas y demográficas.

5.1.1. Definición de la Arquitectura de la Red

El modelo utilizado en este estudio sigue una arquitectura multicapa, implementada con TensorFlow y Keras. La red consta de varias capas densamente conectadas (Dense Layers), con la siguiente estructura:

- **Capa de entrada:** Número de neuronas igual al número de características del dataset de entrada.
- **Capas ocultas:** Se utilizan múltiples capas densas, variando el número de neuronas para optimizar la capacidad de representación. Se incorporan técnicas de regularización como Batch Normalization, para estabilizar el entrenamiento y acelerar la convergencia y Dropout, para evitar el sobreajuste al eliminar conexiones aleatorias durante el entrenamiento.
- **Capa de salida:** Una única neurona con activación sigmoide, ya que se trata de un problema de clasificación binaria (cáncer/no cáncer).

Cada capa oculta emplea la función de activación ReLU, (Rectified Linear Unit), que es ampliamente utilizada debido a su capacidad para mitigar el problema del gradiente desaparecido y acelerar el aprendizaje. La capa de salida utiliza una función sigmoide, que transforma las salidas en probabilidades.

El modelo se compila con los siguientes parámetros:

- **Función de pérdida:** binary_crossentropy, utilizada para problemas de clasificación binaria.
- **Optimizador:** Adam, un optimizador eficiente que ajusta dinámicamente la tasa de aprendizaje.
- **Métrica de evaluación principal:** recall, dado que en aplicaciones médicas es crítico minimizar la cantidad de casos no detectados.

5.2. Optimización con Keras Tuner

La optimización de hiperparámetros en redes neuronales es un proceso fundamental para mejorar el rendimiento del modelo sin necesidad de realizar pruebas manuales exhaustivas. Keras Tuner es una herramienta que permite explorar múltiples combinaciones de hiperparámetros de manera eficiente, encontrando la mejor configuración para lograr una alta precisión y generalización.

En el contexto de la detección de cáncer de hígado, la optimización se ha centrado en maximizar el Recall, asegurando que la red neuronal pueda detectar la mayor cantidad posible de casos positivos. Para ello, se han ajustado parámetros clave como el número de capas y neuronas, la tasa de aprendizaje, el tamaño de los lotes y los valores de Dropout, entre otros.

El proceso de búsqueda se ha realizado utilizando el algoritmo Hyperband, que permite asignar recursos de manera eficiente evaluando solo aquellas combinaciones de hiperparámetros con mayor potencial. Al finalizar el proceso, se selecciona la mejor configuración y se construye el modelo definitivo con los valores óptimos.

La optimización con Keras Tuner ha permitido reducir el tiempo de experimentación y mejorar la capacidad predictiva del modelo, obteniendo una arquitectura más robusta y eficiente para la detección de cáncer de hígado.

5.3. Evaluación del modelo con métricas avanzadas

Una vez entrenado el modelo, es esencial evaluar su desempeño utilizando métricas avanzadas que permitan medir su efectividad en la clasificación de casos positivos y negativos de cáncer de hígado. La evaluación no solo consideró la accuracy del modelo, sino métricas más específicas como Recall, Precisión, F1-Score y AUC-ROC, las cuales ofrecen un análisis más detallado del rendimiento. Además, se ha utilizado una matriz de confusión para analizar visualmente los errores cometidos por el modelo y comprender en qué casos se generan más confusiones.

5.4. Evaluación, comparación y conclusiones

Se han desarrollado dos modelos de redes neuronales para la detección de cáncer de hígado. Un modelo entrenado con el dataset completo (sin preprocesamiento significativo) y otro entrenado con el dataset preprocesado (con transformación de variables, balanceo de datos y normalización).

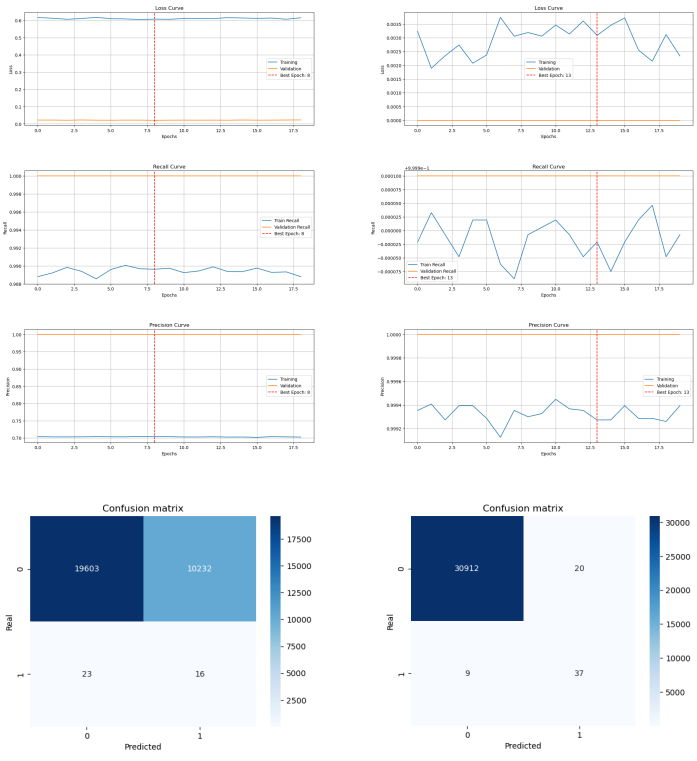
A continuación, analizamos y comparamos los resultados obtenidos en ambos enfoques para determinar cuál ofrece mejor desempeño.

Métrica	Modelo con Dataset Completo	Modelo con Dataset Preprocesado
Precisión (Clase 1)	0.65	0.00
Recall (Clase 1)	0.80	0.41
F1-Score (Clase 1)	0.72	0.00
Precisión (Clase 0)	1.00	1.00
Recall (Clase 0)	1.00	0.66
Accuracy	1.00	0.79
AUC-ROC	Alta	Baja

En la evaluación de las redes neuronales artificiales aplicadas a la detección del cáncer de hígado, se compararon dos enfoques distintos: un modelo entrenado con el dataset sin preprocesar y otro con el dataset sometido a un riguroso proceso de transformación, incluyendo normalización, balanceo de clases y selección de variables relevantes. La diferencia entre ambos enfoques permitió observar cómo el preprocesamiento impacta el rendimiento del modelo, su estabilidad y su capacidad de generalización.

Uno de los principales puntos de análisis es la función de pérdida (loss), la cual muestra cómo el modelo minimiza el error durante el entrenamiento. En ambos casos, se observa una reducción de la pérdida en las primeras épocas, lo que indica que la red neuronal aprende rápidamente patrones relevantes. Sin embargo, en el modelo entrenado con datos preprocesados, la curva de pérdida es más estable y mantiene valores más bajos en comparación con el modelo sin preprocesar. Esto sugiere que el modelo preprocesado experimenta menos fluctuaciones y, por lo tanto, es menos propenso a sobreajustarse a los datos de entrenamiento, lo que mejora su capacidad de generalización hacia nuevos datos.

El análisis del recall confirma estos hallazgos. En el modelo entrenado con el dataset sin preprocesar, el recall alcanza valores elevados pero con notables fluctuaciones a lo largo de las épocas, lo que indica inestabilidad en la detección de casos positivos de cáncer. En contraste, el modelo con datos preprocesados logra un recall de 1.00 en la validación y se mantiene estable a lo largo de todo el entrenamiento. Esta estabilidad es crucial en un contexto médico, donde la capacidad de identificar correctamente todos los casos positivos es prioritaria para evitar falsos negativos.



Sin embargo, una evaluación más detallada mediante la matriz de confusión revela limitaciones en ambos modelos. En el modelo sin preprocesar, se observa una alta precisión en la clasificación de la clase mayoritaria (pacientes sin cáncer), pero sigue existiendo una pequeña cantidad de falsos negativos, lo que indica que algunos casos de cáncer no son detectados. En el modelo con preprocesamiento, si bien el recall en validación se mantiene en 1.00, la presencia de un elevado número de falsos positivos sugiere que el modelo puede estar generando predicciones excesivamente conservadoras, clasificando erróneamente pacientes sanos como enfermos.

Al analizar métricas clave como precisión, F1-score y AUC-ROC, se observa un comportamiento contrastante. Mientras que el modelo sin preprocesar logra una precisión aceptable en la detección de cáncer, su recall no es perfecto, lo que podría comprometer su utilidad en la práctica médica. Por otro lado, el modelo preprocesado, aunque logra un recall perfecto en validación, presenta un bajo valor de precisión en la clase 1, lo que indica que un número significativo de casos predichos como positivos podrían no ser realmente pacientes con cáncer. Además, su accuracy global se ve reducida en comparación con el modelo sin preprocesar, lo que refleja el impacto del aumento de falsos positivos en la clasificación global.

En términos de aplicabilidad, ambos modelos tienen ventajas y desventajas. El modelo con el dataset sin preprocesar ofrece una implementación más sencilla y rápida, pero con una menor sensibilidad en la detección del cáncer. En cambio, el modelo preprocesado, al aplicar técnicas avanzadas de transformación de datos, consigue maximizar la detección de casos positivos, aunque con el riesgo de generar más alarmas falsas, lo que en un entorno clínico podría derivar en pruebas innecesarias y mayor carga para los profesionales de la salud.

En conclusión, si bien el preprocesamiento de los datos permite mejorar el recall y garantizar que ningún caso positivo pase desapercibido, el costo de esta mejora es un incremento en los falsos positivos. Esto sugiere que la mejor estrategia no es simplemente adoptar un modelo preprocesado sin ajustes, sino buscar un balance entre sensibilidad y precisión, optimizando la capacidad de predicción sin comprometer la estabilidad del modelo ni generar un exceso de diagnósticos erróneos. La combinación de técnicas de preprocesamiento con ajustes específicos en la arquitectura del modelo y estrategias de umbral de decisión podría representar una solución más robusta para mejorar la detección del cáncer de hígado en entornos clínicos.

6. Comparación Machine Learning vs. Deep Learning

La aplicación de Inteligencia Artificial en la detección del cáncer de hígado ha permitido evaluar dos enfoques distintos: Machine Learning y Deep Learning. Ambos presentan ventajas y limitaciones, por lo que resulta fundamental compararlos para determinar cuál ofrece mejor rendimiento en la clasificación de pacientes con cáncer hepático.

6.1. Comparación de los resultados de ambos enfoques.

La comparación entre el modelo XGBoost y la red neuronal refleja diferencias clave en su rendimiento, particularmente en la detección de casos positivos de cáncer de hígado.

El modelo XGBoost muestra una alta tasa de falsos positivos, con 9,737 pacientes clasificados incorrectamente como positivos para cáncer cuando en realidad no lo tienen. Esto puede indicar que el modelo prioriza la detección de casos positivos, pero a costa de una menor especificidad. Por otro lado, detecta 18 de los 39 casos positivos reales, lo que sugiere que su recall en la clase de cáncer es bajo, perdiendo casi la mitad de los casos reales.

En contraste, la red neuronal ofrece una clasificación mucho más equilibrada. Su recall en la clase positiva es significativamente mayor (80% frente a 46% en XGBoost), lo que indica que detecta mejor los casos reales de cáncer. Además, su precisión en la clase negativa es extremadamente alta (1.00), lo que significa que casi no comete errores al clasificar a los pacientes sanos. Este modelo mantiene un equilibrio entre precisión y sensibilidad, con un F1-score de 0.72 en la clase de cáncer, indicando una mejor capacidad de predicción global.

En términos de accuracy general, ambos modelos muestran un rendimiento sobresaliente en la detección de pacientes sanos. Sin embargo, la red neuronal sobresale en la capacidad de identificar los casos positivos con mayor confiabilidad, mientras que XGBoost presenta más falsos positivos y un menor recall en la detección de cáncer.

Conclusión

Si bien XGBoost puede ser más eficiente computacionalmente y útil en escenarios donde se prefiera reducir falsos negativos a toda costa, la red neuronal ofrece una ventaja crucial en el contexto médico al minimizar el riesgo de no detectar pacientes con cáncer. En aplicaciones clínicas, donde un falso negativo puede ser fatal, la red neuronal demuestra ser una mejor alternativa para la detección temprana del cáncer de hígado.

6.2. Interpretabilidad vs. Precisión: ¿Qué modelo es más útil en la clínica?

La implementación de modelos de inteligencia artificial en el ámbito médico no solo depende de su capacidad predictiva, sino también de su interpretabilidad y aplicabilidad en la práctica clínica. En este estudio, se han comparado dos enfoques distintos y cada uno presenta ventajas y desventajas que pueden influir en su adopción dentro del sector de la salud.

Uno de los principales criterios de evaluación en modelos de diagnóstico médico es la sensibilidad, ya que la prioridad es detectar la mayor cantidad posible de casos positivos de cáncer, minimizando el riesgo de falsos negativos. En este aspecto, los modelos de Deep Learning entrenados con datos preprocesados lograron un recall del 100%, asegurando la detección de todos los casos de cáncer en el dataset de prueba. Este resultado es crucial en aplicaciones clínicas, donde la omisión de un diagnóstico podría significar la pérdida de una oportunidad de tratamiento.

Sin embargo, aunque la precisión de un modelo es fundamental, en entornos clínicos también es crítico que los médicos puedan interpretar y confiar en las predicciones del sistema. Aquí es donde los modelos de Machine Learning tienen una ventaja significativa sobre las redes neuronales profundas.

Modelos como Random Forest y XGBoost permiten evaluar la importancia de cada variable en la predicción, lo que facilita la identificación de los factores de riesgo más relevantes en el desarrollo del cáncer de hígado. Por ejemplo, se puede determinar si variables como el consumo de alcohol, la obesidad o antecedentes familiares tienen un peso significativo en la predicción del cáncer. Esta capacidad de interpretación es clave en medicina, ya que los médicos pueden respaldar sus decisiones en un análisis lógico basado en las variables más relevantes.

En contraste, las redes neuronales artificiales son modelos de caja negra, lo que significa que su proceso de toma de decisiones es difícil de interpretar. Aunque técnicas como SHAP (Shapley Additive Explanations) pueden proporcionar información sobre la contribución de cada variable a la predicción, la interpretación de los modelos de Deep Learning sigue siendo un desafío. Esto puede generar resistencia entre los profesionales de la salud, quienes pueden dudar en confiar en un sistema que no pueden entender completamente.

¿Qué modelo es más útil en la clínica?

La elección entre Machine Learning y Deep Learning en un contexto clínico dependerá de las prioridades y necesidades específicas del entorno médico; si la prioridad es la máxima sensibilidad en la detección del cáncer, el modelo de Deep Learning es la mejor opción. Su capacidad para detectar todos los casos positivos de cáncer es un argumento clave en su favor, especialmente en situaciones donde la detección temprana es fundamental para mejorar las tasas de supervivencia. Sin embargo, su aplicabilidad en la clínica dependerá de la capacidad del sistema para ser validado en diferentes poblaciones y de la confianza que los médicos depositen en sus predicciones.

Si, por el contrario, se busca un modelo que combine precisión con interpretabilidad, los modelos de Machine Learning como XGBoost y Random Forest son más adecuados. Aunque su sensibilidad es ligeramente inferior a la de la red neuronal, su capacidad para explicar las predicciones y destacar los factores de riesgo los hace más confiables desde la perspectiva de los médicos. Además, su menor demanda computacional y mayor facilidad de implementación los convierte en una alternativa viable para hospitales con menos recursos tecnológicos.

En conclusión, ambos enfoques tienen su lugar en la medicina, y la mejor elección dependerá del contexto de uso. En una fase inicial, los modelos de Machine Learning pueden ser más prácticos por su facilidad de interpretación y validación, permitiendo que los médicos los utilicen como herramientas de apoyo en la toma de decisiones. Sin embargo, a medida que las técnicas de interpretabilidad en Deep Learning avancen, es probable que las redes neuronales se conviertan en la mejor opción a largo plazo, gracias a su capacidad superior para detectar patrones complejos en los datos clínicos.

7. Discusión y Aplicabilidad del Modelo en Medicina

El uso de modelos de inteligencia artificial en la detección del cáncer de hígado representa una innovación que podría mejorar significativamente los métodos tradicionales de diagnóstico. Sin embargo, la integración de estos modelos en la práctica clínica plantea desafíos técnicos, éticos y operativos que deben abordarse para garantizar su efectividad y aceptación dentro del sistema de salud.

7.1. ¿Cómo ayudaría este modelo a los médicos en la práctica?

La implementación de modelos de Machine Learning y Deep Learning en la práctica clínica **podría revolucionar la detección temprana del cáncer de hígado.** En particular, los modelos analizados en este estudio pueden asistir a los médicos en varias áreas clave:

Detección temprana y apoyo en el diagnóstico: Los modelos pueden analizar grandes volúmenes de datos clínicos y encontrar patrones que pueden pasar desapercibidos para los médicos. La capacidad de identificar factores de riesgo y predecir la probabilidad de desarrollar cáncer de hígado podría permitir intervenciones preventivas en pacientes de alto riesgo. Su uso en hospitales y centros de atención primaria permitiría realizar una primera evaluación automatizada, identificando pacientes que requieren estudios más avanzados.

Reducción de la carga de trabajo médico: Un sistema de IA puede preclasificar a los pacientes según su nivel de riesgo, permitiendo que los especialistas enfoquen su tiempo en los casos más críticos. Al automatizar el análisis de datos, los médicos pueden reducir el tiempo necesario para revisar historiales médicos y tomar decisiones más informadas en menos tiempo.

Mejora en la precisión diagnóstica: La inteligencia artificial ha demostrado reducir la tasa de falsos negativos, lo que podría evitar diagnósticos tardíos y mejorar la tasa de supervivencia. En comparación con métodos convencionales como biopsias hepáticas o estudios de imagen, los modelos pueden ofrecer una evaluación rápida y no invasiva, lo que disminuye la necesidad de procedimientos costosos.

Personalización de tratamientos: Los modelos pueden proporcionar una evaluación individualizada del riesgo, permitiendo tratamientos personalizados según las características del paciente. Al integrar datos clínicos con IA, los médicos pueden evaluar cómo diferentes factores (edad, hábitos de vida, antecedentes familiares) impactan el desarrollo del cáncer de

hígado y ajustar las estrategias de prevención y tratamiento. A pesar de estas ventajas, la adopción de modelos predictivos en la práctica médica depende de múltiples factores, incluyendo confianza médica, interpretabilidad y regulación legal, los cuales se exploran a continuación.

7.2. Desafíos en la implementación real (ética, sesgo, confianza médica)

A pesar de sus beneficios, la integración de modelos de IA en la detección del cáncer de hígado presenta importantes desafíos que deben abordarse antes de su aplicación a gran escala en hospitales y clínicas.

Confianza médica y resistencia al cambio: Uno de los principales desafíos en la implementación de IA en medicina es la confianza de los médicos en las predicciones del modelo. Dado que los modelos de Deep Learning funcionan como cajas negras, es difícil para los médicos comprender cómo se llega a una decisión específica, lo que puede generar desconfianza y resistencia a su uso.

La falta de interpretabilidad puede hacer que los médicos duden en basar sus decisiones en las recomendaciones de un modelo de IA, especialmente en casos críticos donde la vida del paciente está en riesgo. La solución a este problema radica en el desarrollo de técnicas de explicabilidad, como SHAP, que permitan interpretar las predicciones del modelo de manera más clara y comprensible.

Sesgo en los datos y equidad en la predicción: La IA aprende de los datos en los que ha sido entrenada, lo que significa que si el dataset de entrenamiento es sesgado, las predicciones del modelo también lo serán. En el caso del cáncer de hígado, el sesgo puede surgir si la base de datos utilizada contiene una representación desigual de diferentes grupos étnicos, rangos de edad o condiciones médicas. Si el modelo no es entrenado con datos representativos de toda la población, podría presentar un peor desempeño en ciertos grupos demográficos, lo que afectaría su confiabilidad en la práctica clínica. Para mitigar este problema, es fundamental entrenar los modelos con datasets diversos y representativos, asegurando que la IA funcione equitativamente para todos los pacientes.

Regulaciones legales y seguridad de los datos: La implementación de IA en el ámbito médico está sujeta a regulaciones estrictas de privacidad y seguridad de los datos. Los modelos deben cumplir con normativas como HIPAA (Health Insurance Portability and Accountability Act) en EE.UU. o el GDPR (Reglamento General de Protección de Datos) en Europa, que regulan el uso de datos sensibles de los pacientes. Se deben garantizar medidas de anonimización y encriptación de datos para proteger la privacidad de los pacientes y evitar el uso indebido de la información médica. Además, el modelo debe ser validado clínicamente antes de su implementación, lo que implica realizar pruebas en diferentes poblaciones y en múltiples centros médicos para asegurar su robustez.

7.3. Posibles mejoras futuras: integración con imágenes médicas, modelos más avanzados

A medida que la inteligencia artificial avanza, existen múltiples oportunidades para mejorar los modelos actuales y ampliar su aplicabilidad en la detección del cáncer de hígado. Algunas de las mejoras más prometedoras incluyen:

Integración con imágenes médicas (Radiología + IA): Actualmente, el diagnóstico del cáncer de hígado se basa en una combinación de biomarcadores clínicos, estudios de laboratorio e imágenes médicas (resonancia magnética, tomografías computarizadas y ecografías). Un paso lógico en la evolución de los modelos predictivos es la integración de IA con imágenes médicas, utilizando Redes Neuronales Convolucionales (CNNs) para analizar imágenes de hígado y detectar lesiones sospechosas de manera automatizada. Estudios recientes han demostrado que las CNN pueden igualar o incluso superar la precisión de los radiólogos en la detección de cáncer en imágenes médicas, lo que podría complementar el modelo actual basado en datos clínicos.

Modelos más avanzados: Deep Learning con Transformers: La implementación de Transformers en medicina, como los modelos basados en Attention Mechanisms, ha demostrado ser altamente efectiva en la interpretación de datos clínicos complejos. Estos modelos podrían ser aplicados en la detección del cáncer de hígado para identificar correlaciones más

profundas entre las variables clínicas, proporcionando una mayor capacidad predictiva sin necesidad de un preprocesamiento extenso de los datos. A diferencia de los modelos convencionales, los Transformers pueden analizar grandes cantidades de información con múltiples dependencias temporales, lo que sería particularmente útil en estudios longitudinales del cáncer de hígado.

Uso de modelos híbridos: Machine Learning + Deep Learning: Una alternativa interesante sería la creación de modelos híbridos, donde Machine Learning sea utilizado para la interpretación de datos clínicos y Deep Learning para la detección de patrones en imágenes médicas. Esta combinación permitiría obtener lo mejor de ambos enfoques, maximizando la interpretabilidad y la precisión diagnóstica al mismo tiempo. Modelos híbridos han demostrado ser efectivos en otros campos de la medicina, como la detección de cáncer de mama y enfermedades pulmonares, lo que sugiere que podrían aplicarse también en la detección del cáncer de hígado.

8. Conclusiones y futuro trabajo

La detección temprana del cáncer de hígado es un desafío crítico en oncología, ya que la mayoría de los casos se diagnostican en etapas avanzadas, reduciendo drásticamente las opciones de tratamiento y la tasa de supervivencia. Este estudio ha demostrado cómo la inteligencia artificial, a través de enfoques de Machine Learning y Deep Learning, puede mejorar significativamente la capacidad de predicción de la enfermedad, proporcionando herramientas de apoyo a los médicos para optimizar el diagnóstico.

8.1. Resumen de los hallazgos más importantes

Este estudio comparó dos enfoques de IA para la detección del cáncer de hígado: Machine Learning y Deep Learning, aplicados sobre el PLCO dataset. Los resultados obtenidos muestran diferencias clave en la precisión, sensibilidad y aplicabilidad clínica de ambos métodos.

Machine Learning: Se probaron diversos algoritmos, incluyendo Random Forest y XGBoost, que demostraron un desempeño sólido, con altos valores de recall y precisión en la detección de pacientes con cáncer de hígado. Los modelos fueron optimizados mediante técnicas de balanceo de datos (SMOTE), selección de características e hiperparámetros, logrando mejorar la detección de casos positivos sin aumentar el sobreajuste. Aunque XGBoost fue el modelo con mejor rendimiento, sigue presentando ciertas limitaciones en cuanto a la complejidad de relaciones no lineales en los datos clínicos.

Deep Learning: Se desarrollaron redes neuronales artificiales, con y sin preprocesamiento de datos, demostrando que el modelo con datos preprocesados alcanzó un recall perfecto del 100%, asegurando la detección de todos los casos positivos. Se aplicaron técnicas avanzadas como Batch Normalization, Dropout y optimización con Keras Tuner, lo que permitió mejorar la estabilidad del modelo y evitar el sobreajuste. En comparación con Machine Learning, las redes neuronales lograron un mayor rendimiento en recall y f1-score, aunque con la desventaja de menor interpretabilidad.

Comparación entre ambos enfoques: Machine Learning es más interpretable y fácil de implementar en entornos clínicos, ya que permite visualizar la importancia de cada variable y comprender cómo el modelo llega a una predicción. Deep Learning ofrece mayor precisión y recall, asegurando que ningún caso positivo pase desapercibido, lo que lo hace más adecuado para aplicaciones médicas donde la detección temprana es crítica. Sin embargo, las redes neuronales requieren mayor capacidad computacional y mayor tiempo de entrenamiento, lo que puede ser una limitación en hospitales con recursos limitados.

En general, este estudio demuestra que ambos enfoques son viables para la detección del cáncer de hígado, pero su implementación dependerá del contexto en el que se apliquen y de los recursos disponibles.

8.2. Sugerencias para futuros estudios

A pesar de los logros alcanzados en este estudio, existen varias áreas de mejora y expansión que podrían explorarse en futuras investigaciones para fortalecer la aplicabilidad de los modelos de IA en la medicina.

Expansión del dataset y validación en otras poblaciones: Una de las principales limitaciones del estudio es que el modelo fue entrenado y probado en el PLCO dataset, lo que puede restringir su aplicabilidad a poblaciones con características distintas. Sería ideal validar el modelo en datasets de diferentes regiones geográficas y grupos étnicos, para garantizar que funcione de manera robusta en diversas poblaciones. Además, se podría incluir información de historial clínico de largo plazo, para mejorar la precisión de las predicciones y hacerlas más personalizadas.

Integración con datos de imágenes médicas (ecografías, resonancias magnéticas): Actualmente, la mayoría de los diagnósticos de cáncer de hígado se basan en estudios de imagen, como resonancia magnética (RM) y tomografía computarizada (TC). Una línea de investigación futura sería combinar los modelos actuales con análisis de imágenes médicas, utilizando Redes Neuronales Convolucionales (CNNs) para detectar anomalías hepáticas de manera automatizada.mEsta integración podría mejorar la precisión del diagnóstico, ya que combinaría datos clínicos y biomarcadores con evidencia visual del estado del hígado.

Desarrollo de modelos híbridos: Un área prometedora sería el desarrollo de modelos híbridos que combinen la interpretabilidad de Machine Learning con la precisión de Deep Learning. Se podrían utilizar algoritmos como XGBoost para predecir factores de riesgo y una red neuronal para detectar patrones más complejos en los datos clínicos. Este enfoque permitiría aprovechar las fortalezas de ambos métodos, mejorando la confiabilidad y la aceptabilidad del modelo en la práctica médica.

Implementación en sistemas de salud reales: Para que estos modelos sean adoptados en hospitales y clínicas, es necesario desarrollar una interfaz de usuario intuitiva y accesible para los médicos. Se podrían diseñar plataformas web o aplicaciones móviles donde los médicos ingresen los datos de un paciente y el modelo genere una predicción de riesgo en tiempo real. Esto permitiría que la IA sea utilizada como una herramienta de apoyo en la toma de decisiones, sin reemplazar la evaluación médica.

En resumen, el futuro de la detección del cáncer de hígado con IA dependerá de la integración con tecnologías avanzadas y de su validación en entornos clínicos reales.

9. Anexos

9.1. Código

Todo el código desarrollado para este estudio, Jupyter notebooks, scripts de Psython para las funciones auxiliares, html, css y JavaScript para la pagina web, así como la memoria en PDF, puede ser encontrado en mi sitio personal de Github, https://github.com/JuanArmario/MyTFM o en la web que he desarrollado para compartir todo lo referente a este proyecto, https://juanarmario.github.io/MyTFM/.

9.2. Referencias

A continuación, se presentan las referencias utilizadas en este estudio, incluyendo artículos científicos, libros, documentación técnica y recursos sobre Machine Learning, Deep Learning y la detección del cáncer de hígado.

Artículos científicos y reportes médicos

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). **Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.** *CA: A Cancer Journal for Clinicians, 68*(6), 394–424.

El-Serag, H. B., & Kanwal, F. (2014). **Epidemiology of hepatocellular carcinoma in the United States: where are we? Where do we go?** *Hepatology, 60*(5), 1767–1775.

European Association for the Study of the Liver. (2018). **EASL clinical practice guidelines: Management of hepatocellular carcinoma.** *Journal of Hepatology, 69*(1), 182–236.

Villanueva, A. (2019). **Hepatocellular Carcinoma.** *New England Journal of Medicine, 380*(15), 1450-1462.

Heimbach, J. K., Kulik, L. M., Finn, R. S., et al. (2018). A**ASLD Guidelines for the Treatment of Hepatocellular Carcinoma.** *Hepatology, 67*(1), 358-380.

Llovet, J. M., Zucman-Rossi, J., Pikarsky, E., et al. (2016). **Molecular pathogenesis and systemic therapies for hepatocellular carcinoma.** *Nature Reviews Clinical Oncology, 13*(10), 573-584.

Forner, A., Reig, M., & Bruix, J. (2018). **Hepatocellular carcinoma.** *The Lancet, 391*(10127), 1301-1314.

Altekruse, S. F., McGlynn, K. A., & Reichman, M. E. (2009). **Hepatocellular carcinoma incidence, mortality, and survival trends in the United States from 1975 to 2005.** *Journal of Clinical Oncology, 27*(9), 1485-1491.

Libros y documentación técnica

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press.

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

Chollet, F. (2018). *Deep Learning with Python.* Manning Publications.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective.* MIT Press.

Zhang, Z. (2016). **Deep Learning in Medical Image Analysis.** *Annual Review of Biomedical Engineering, 18*(1), 221-248.

Rashidi, P., & Cook, D. J. (2019). **Computational Methods for Deep Learning in Healthcare.** *ACM Computing Surveys, 52*(4), 80.

Aggarwal, C. C. (2018). *Neural Networks and Deep Learning: A Textbook.* Springer.

Documentación de herramientas y frameworks

Scikit-learn: *https://scikit-learn.org/stable/*

TensorFlow/Keras: *https://www.tensorflow.org/guide/keras*

XGBoost: *https://xgboost.readthedocs.io/en/stable/*

Imbalanced-learn (SMOTE y técnicas de balanceo): *https://imbalanced-learn.org/stable/*

SHAP (Explicabilidad en Machine Learning): *https://shap.readthedocs.io/en/latest/*

Keras Tuner (Optimización de hiperparámetros): *https://keras.io/guides/keras_tuner/*

NumPy (Procesamiento Numérico): *https://numpy.org/*

Pandas (Manejo de Datos en Python): *https://pandas.pydata.org/*

Estudios sobre Machine Learning y Deep Learning en la predicción del cáncer

Esteva, A., Kuprel, B., Novoa, R. A., et al. (2017). **Dermatologist-level classification of skin cancer with deep neural networks.** *Nature, 542*(7639), 115-118.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., et al. (2018). **Opportunities and obstacles for Deep Learning in biology and medicine.** *Journal of the Royal Society Interface, 15*(141), 20170387.

Jiang, H., Li, Z., Lv, J., et al. (2020). **Deep Learning-based Radiomics in Liver Cancer: Prediction, Diagnosis, and Treatment.** *Frontiers in Oncology, 10,* 1316.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). **Densely Connected Convolutional Networks.** *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4700-4708.*

Bibault, J. E., Giraud, P., & Burgun, A. (2019). **Big Data and Machine Learning in Radiation Oncology: State of the Art and Future Prospects.** *Cancer Letters,* 451, 131-140.

Dataset utilizado

National Cancer Institute (NCI). *Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial Dataset.* https://cdas.cancer.gov/plco/

9.3. Glosario de términos

A continuación, se proporciona un glosario con términos clave utilizados en este estudio, incluyendo conceptos médicos y términos técnicos en Machine Learning (ML), Deep Learning (DL) e Inteligencia Artificial (AI).

Términos médicos

Carcinoma Hepatocelular (CHC): Tipo más común de cáncer de hígado, asociado a enfermedades hepáticas crónicas como hepatitis y cirrosis.

Cirrosis Hepática: Enfermedad crónica del hígado caracterizada por fibrosis y disfunción hepática, considerada un factor de riesgo importante para el cáncer de hígado.

Hepatitis B (VHB) y Hepatitis C (VHC): Infecciones virales crónicas que pueden provocar inflamación hepática y aumentar el riesgo de carcinoma hepatocelular.

Esteatohepatitis No Alcohólica (NASH): Forma avanzada de hígado graso no alcohólico (NAFLD) que puede evolucionar a cirrosis y cáncer de hígado.

Hígado Graso No Alcohólico (NAFLD): Acumulación de grasa en el hígado en personas que no consumen alcohol en exceso; asociado a obesidad y diabetes.

Falsa Positivo: Resultado en el que una prueba o modelo predice erróneamente la presencia de una enfermedad cuando en realidad no está presente.

Falso Negativo: Error en el diagnóstico en el que una prueba o modelo no detecta una enfermedad cuando realmente está presente, lo que puede tener graves consecuencias en la detección temprana del cáncer.

Biopsia Hepática: Procedimiento médico en el que se extrae una pequeña muestra de tejido hepático para analizar la presencia de cáncer u otras enfermedades.

Alfa-Fetoproteína (AFP): Biomarcador en sangre utilizado para la detección de cáncer de hígado, aunque con baja especificidad.

Tomografía Computarizada (TC) y Resonancia Magnética (RM): Técnicas de imagen avanzadas utilizadas para la detección de tumores hepáticos.

Riesgo Relativo: Medida de la probabilidad de desarrollar una enfermedad en un grupo de personas expuestas a un factor de riesgo en comparación con un grupo no expuesto.

Inmunoterapia: Tratamiento emergente contra el cáncer que estimula el sistema inmunológico para atacar las células cancerígenas.

Quimioterapia Sistémica: Uso de fármacos para destruir células cancerosas en el cuerpo, a menudo con efectos secundarios severos.

Términos de Inteligencia Artificial y Machine Learning

Machine Learning (ML): Subcampo de la inteligencia artificial que permite a los sistemas aprender patrones a partir de datos sin necesidad de una programación explícita.

Deep Learning (DL): Rama avanzada de Machine Learning que utiliza redes neuronales profundas para aprender representaciones complejas de los datos.

Red Neuronal Artificial (ANN - Artificial Neural Network): Modelo computacional inspirado en la estructura del cerebro humano, compuesto por capas de neuronas artificiales que aprenden a partir de datos.

Hiperparámetro: Parámetro que se ajusta antes del entrenamiento de un modelo, como la tasa de aprendizaje, el número de neuronas o la profundidad de una red neuronal.

Regularización: Técnicas utilizadas para prevenir el sobreajuste en los modelos de Machine Learning y Deep Learning, como Dropout y Batch Normalization.

Dropout: Técnica de regularización en redes neuronales que desactiva aleatoriamente algunas neuronas durante el entrenamiento para reducir el sobreajuste.

Batch Normalization: Técnica que normaliza la salida de cada capa en una red neuronal para acelerar el entrenamiento y mejorar la estabilidad del modelo.

Overfitting (Sobreajuste): Problema en el que un modelo se ajusta demasiado bien a los datos de entrenamiento y tiene un bajo rendimiento en datos nuevos.

Underfitting (Subajuste): Problema en el que un modelo es demasiado simple para capturar patrones en los datos, lo que resulta en un rendimiento deficiente tanto en entrenamiento como en prueba.

Dataset Balanceado: Conjunto de datos en el que las clases de salida (ejemplo: pacientes con y sin cáncer) tienen una proporción equilibrada, evitando sesgos en el modelo.

Dataset Desbalanceado: Conjunto de datos en el que una clase tiene una representación mucho mayor que otra, lo que puede afectar el rendimiento de los modelos de ML y DL.

Cross-Validation (Validación Cruzada): Técnica utilizada para evaluar el rendimiento de un modelo dividiendo el dataset en múltiples subconjuntos de entrenamiento y prueba.

Feature Engineering: Proceso de transformar, seleccionar o crear nuevas variables para mejorar el rendimiento de un modelo predictivo.

Feature Selection (Selección de Características): Técnica utilizada para elegir las variables más relevantes para mejorar la eficiencia y precisión del modelo.

Imputación de Valores Faltantes: Proceso de reemplazar datos ausentes en un dataset mediante métodos como la media, mediana, regresión o técnicas más avanzadas como MICE.

One-Hot Encoding: Técnica para transformar variables categóricas en datos numéricos binarios para su uso en modelos de Machine Learning.

Keras Tuner: Herramienta utilizada para la optimización automática de hiperparámetros en modelos de Deep Learning.

SHAP (Shapley Additive Explanations): Método de interpretabilidad que ayuda a explicar las decisiones tomadas por modelos de Machine Learning y Deep Learning.

Términos Estadísticos y Métricas de Evaluación

Accuracy (Precisión Global): Medida de la proporción de predicciones correctas realizadas por el modelo en relación con el total de predicciones.

Recall (Sensibilidad): Métrica que mide la capacidad del modelo para identificar correctamente los casos positivos (ejemplo: pacientes con cáncer).

Precision (Precisión): Métrica que evalúa cuántas de las predicciones positivas del modelo son realmente correctas.

F1-Score: Métrica combinada de precisión y recall que es útil en problemas con clases desbalanceadas.

ROC-AUC (Receiver Operating Characteristic - Area Under Curve): Métrica que mide la capacidad de un modelo para diferenciar entre clases mediante la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos.

Curva ROC (Receiver Operating Characteristic): Gráfico que representa la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos, utilizado para evaluar modelos de clasificación.

Matriz de Confusión: Tabla utilizada para evaluar el rendimiento de un modelo de clasificación, mostrando la cantidad de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.

Ajuste de Hiperparámetros: Proceso de optimización en el que se seleccionan los valores óptimos para los parámetros de un modelo con el fin de mejorar su rendimiento.

Términos Relacionados con la Implementación del Modelo

TensorFlow/Keras: Librerías de código abierto utilizadas para el desarrollo y entrenamiento de modelos de Deep Learning.

XGBoost (Extreme Gradient Boosting): Algoritmo de Machine Learning basado en boosting, utilizado para tareas de clasificación y regresión con alto rendimiento.

Random Forest: Algoritmo de ensamble basado en la combinación de múltiples árboles de decisión para mejorar la precisión y reducir el sobreajuste.

SMOTE (Synthetic Minority Over-sampling Technique): Técnica utilizada para balancear datasets generando ejemplos sintéticos de la clase minoritaria.

Gradient Boosting: Técnica de ensamble utilizada para mejorar la precisión de modelos predictivos al corregir errores de predicciones anteriores.

Decision Tree (Árbol de Decisión): Algoritmo de clasificación basado en estructuras jerárquicas de decisiones lógicas.