UNIVERSIDAD DE GUADALAJARA

CENTRO UNIVERSITARIO DE CIENCIAS ECONÓMICO ADMINISTRATIVAS

COORDINACIÓN DE POSGRADO

MAESTRÍA EN CIENCIA DE LOS DATOS

Detección temprana de enfermedades cerebrovasculares a través del análisis de imágenes de fondo de ojo

P R E S E N T A

MARÍA PALOMA PRADO DURÁN

Fecha: MAYO 2025

CUCEA
El mejor lugar para el talento

MCD MAESTRÍA EN CIENCIA DE LOS DATOS

Contenido

INTRODUCCIÓN 2
MODELOS 3
PROPUESTA DE APLICACIÓN DE MODELOS 29
ESTRUCTURA DEL TRABAJO EN PYTHON 31
CONCLUSIÓN 31
REFERENCIAS 32

INTRODUCCIÓN

En el marco del desarrollo de soluciones basadas en inteligencia artificial para el ámbito de la salud, el presente trabajo tiene como objetivo analizar diversos modelos de aprendizaje automático, tanto supervisados como no supervisados, con el fin de identificar sus ventajas y limitaciones para su posible aplicación en un proyecto de detección temprana de enfermedades cerebrovasculares mediante el análisis de imágenes de fondo de ojo. Este tipo de imágenes ofrece una vía no invasiva y accesible para evaluar indirectamente la salud vascular cerebral, dada la conexión entre la microvasculatura retiniana y el sistema circulatorio cerebral.

El análisis se centra en modelos comúnmente utilizados en la ciencia de los datos, como redes neuronales, máquinas de soporte vectorial, árboles de decisión, métodos de ensamblado como Random Forest y XGBoost, así como técnicas de reducción de dimensionalidad y agrupamiento, tales como PCA, t-SNE y autoencoders. A través de la evaluación comparativa de sus capacidades predictivas, requerimientos computacionales, interpretabilidad y adaptabilidad al contexto médico, se busca establecer una base sólida para seleccionar los modelos más adecuados en función de los requerimientos técnicos y clínicos del proyecto.

Este estudio no solo facilitará la toma de decisiones en las etapas posteriores del diseño del sistema de diagnóstico, sino que también contribuirá a fundamentar científicamente el enfoque metodológico adoptado, considerando el equilibrio entre rendimiento, explicabilidad y viabilidad práctica en un entorno de aplicación

médica crítica.

MODELOS

Comenzaremos con los *Modelos de Aprendizaje Supervisado*, los cuales como lo menciona IBM (s.f.-b) es “también conocido como aprendizaje automático supervisado, es una subcategoría del aprendizaje automático y la inteligencia artificial. Se define por su uso de conjuntos de datos etiquetados para entrenar algoritmos que clasifiquen datos o predigan resultados con precisión.” Es decir, “a medida que los datos de entrada se introducen en el modelo, este ajusta sus ponderaciones hasta que el modelo se haya adaptado adecuadamente, lo que ocurre como parte del proceso de validación cruzada.”

Como primer modelo tenemos el de ***Regresión Lineal***, el cual, es una técnica estadística que modela la relación entre una variable dependiente y una o más variables independientes, asumiendo una relación lineal entre ellas. En el contexto de la detección temprana de enfermedades cerebrovasculares mediante el análisis de imágenes de fondo de ojo, su aplicación presenta ciertas ventajas y desventajas, que se exponen a continuación.

Para el caso de las **ventajas**, tenemos:

- ☑ SIMPLICIDAD E INTERPRETABILIDAD: La regresión lineal es fácil de entender e interpretar, lo que facilita la comunicación de los resultados a profesionales clínicos y no técnicos.
- ☑ EFICIENCIA COMPUTACIONAL: Requiere menos recursos computacionales en comparación con modelos más complejos, lo que permite una implementación rápida y eficiente.
- ☑ APLICACIÓN EN ESTUDIOS MÉDICOS: Se ha utilizado en medicina para predecir riesgos de enfermedades en función de factores de riesgo conocidos, lo que demuestra su utilidad en contextos clínicos. *(Podrez, 2023)*

Pero desafortunadamente, también se tienen **desventajas** al aplicarlo:

- ☒ SUPOSICIÓN DE LINEALIDAD: Asume una relación lineal entre las variables, lo que puede no ser adecuado para datos complejos como imágenes

médicas, donde las relaciones pueden ser no lineales.

- ☒ SENSIBILIDAD A MULTICOLINEALIDAD: La presencia de variables independientes altamente correlacionadas puede afectar la estabilidad y precisión de los coeficientes estimados. (Godwa & Shorck, s.f.)
- ☒ LIMITACIONES EN LA PREDICCIÓN DE VARIABLES CATEGÓRICAS: No es adecuada para predecir variables categóricas sin modificaciones, lo que limita su aplicabilidad en ciertos escenarios clínicos. (Guyatt, et al, 1995)

Ante esto podemos decir que, aunque la regresión lineal ofrece ventajas en términos de simplicidad y eficiencia, sus limitaciones en capturar relaciones no lineales y suposiciones estrictas pueden restringir su aplicabilidad en el análisis de imágenes de fondo de ojo para la detección de enfermedades cerebrovasculares.

Ahora pasamos al modelo de ***Regresión Logística***, que “es una técnica estadística ampliamente utilizada para modelar la probabilidad de ocurrencia de un evento binario, como la presencia o ausencia de una enfermedad. En el contexto de la detección temprana de enfermedades cerebrovasculares mediante el análisis de imágenes de fondo de ojo, la regresión logística puede ser una herramienta valiosa”. A continuación, se presentan sus principales ventajas y desventajas:

Para el caso de las **ventajas**, tenemos que como se menciona en SimpliRoute (2023), se destaca:

- ☑ INTERPRETABILIDAD Y SIMPLICIDAD: La regresión logística es fácil de implementar e interpretar, lo que permite a los profesionales de la salud comprender cómo las variables independientes influyen en la probabilidad de un evento, facilitando la toma de decisiones clínicas.
- ☑ EFICIENCIA COMPUTACIONAL: Requiere menos recursos computacionales en comparación con modelos más complejos, lo que permite su aplicación en sistemas con capacidades limitadas.
- ☑ MANEJO DE VARIABLES CATEGÓRICAS: Puede manejar tanto variables independientes continuas como categóricas, lo que la hace versátil para diferentes tipos de datos clínicos.

- ☑ Estimación de probabilidades: Proporciona probabilidades asociadas a cada clase, lo que es útil para evaluar el riesgo de un paciente y tomar decisiones basadas en umbrales específicos.

Pero desafortunadamente, también se tienen **desventajas** al aplicarlo (GeeksforGeeks, 2023):

- ☒ Suposición de linealidad: Asume una relación lineal entre las variables independientes y el logaritmo de las probabilidades, lo que puede no ser adecuado para datos complejos como las imágenes médicas.
- ☒ Limitaciones con datos no linealmente separables: La regresión logística puede tener dificultades para clasificar correctamente los datos que no son linealmente separables, lo que puede afectar su rendimiento en problemas complejos.
- ☒ Sensibilidad a variables irrelevantes: La inclusión de variables independientes irrelevantes puede afectar negativamente el rendimiento del modelo, requiriendo una cuidadosa selección de características.
- ☒ Rendimiento inferior en comparación con modelos más complejos: En problemas donde las relaciones entre variables son altamente no lineales, modelos más complejos como las redes neuronales pueden superar a la regresión logística en términos de precisión.

En sumo podemos ver que la regresión logística es una herramienta útil para la clasificación binaria y ofrece ventajas significativas en términos de interpretabilidad y eficiencia. Sin embargo, en el contexto del análisis de imágenes de fondo de ojo para la detección de enfermedades cerebrovasculares, donde las relaciones entre variables pueden ser complejas y no lineales, es posible que modelos más avanzados proporcionen un rendimiento superior. Por lo tanto, la regresión logística puede servir como un modelo base o de referencia, pero se recomienda explorar modelos más sofisticados para mejorar la precisión diagnóstica.

En cuanto a los *Árboles de Decisión*, tal y como se menciona en IBM (s.f.-c), estos "son algoritmos de aprendizaje supervisado que se utilizan tanto para tareas de clasificación como de regresión. Su estructura jerárquica permite dividir los datos en subconjuntos basados en características específicas, facilitando la toma de decisiones". A continuación, se presentan las principales ventajas y desventajas de aplicar árboles de decisión en el contexto de la detección temprana de enfermedades cerebrovasculares mediante el análisis de imágenes de fondo de ojo.

Para el caso de las **ventajas**, tenemos que como se menciona en FasterCapital (2025), se destaca:

- ☑ INTERPRETABILIDAD Y TRANSPARENCIA: Los árboles de decisión son fáciles de entender e interpretar, lo que permite a los profesionales de la salud comprender cómo se llega a una conclusión diagnóstica. Esta característica es especialmente valiosa en el ámbito médico, donde la explicabilidad es crucial.
- ☑ MANEJO DE DIFERENTES TIPOS DE DATOS: Pueden manejar tanto variables numéricas como categóricas, lo que los hace versátiles para diversos tipos de datos clínicos.
- ☑ NO REQUIEREN SUPOSICIONES SOBRE LA DISTRIBUCIÓN DE LOS DATOS: Al ser modelos no paramétricos, no hacen suposiciones sobre la distribución de los datos, lo que les permite adaptarse a diferentes conjuntos de datos sin necesidad de transformaciones complejas.
- ☑ EFICIENCIA COMPUTACIONAL: Son relativamente rápidos de entrenar y ejecutar, lo que es beneficioso cuando se trabaja con grandes volúmenes de datos, como imágenes médicas de alta resolución.

Pero desafortunadamente, también se tienen **desventajas** al aplicarlo (Ultralytics, s.f.):

- ☒ PROPENSIÓN AL SOBREAJUSTE: Los árboles de decisión pueden volverse demasiado complejos y ajustarse demasiado a los datos de entrenamiento, capturando ruido en lugar de patrones significativos. Esto puede llevar a un rendimiento deficiente en datos no vistos.

- ☒ INESTABILIDAD: Pequeños cambios en los datos de entrenamiento pueden resultar en estructuras de árbol significativamente diferentes, lo que afecta la consistencia del modelo.
- ☒ SESGO HACIA CARACTERÍSTICAS CON MÁS NIVELES: Los árboles de decisión pueden estar sesgados hacia características con más niveles o clases dominantes, lo que puede ser problemático en conjuntos de datos desequilibrados.
- ☒ LIMITACIONES EN LA CAPTURA DE RELACIONES COMPLEJAS: Pueden tener dificultades para capturar relaciones complejas entre variables, especialmente en datos de alta dimensionalidad como las imágenes médicas.

Ante esto, observamos que los árboles de decisión ofrecen una herramienta valiosa para la detección temprana de enfermedades cerebrovasculares mediante el análisis de imágenes de fondo de ojo, gracias a su interpretabilidad y eficiencia. Sin embargo, es importante ser consciente de sus limitaciones, como la propensión al sobreajuste y la inestabilidad. En aplicaciones clínicas, donde la precisión y la consistencia son fundamentales, se recomienda considerar técnicas complementarias o modelos más avanzados que puedan mitigar estas desventajas.

Siguiendo con los modelos tenemos *Máquinas de Soporte Vectorial (SVM)*, que son algoritmos de aprendizaje supervisado ampliamente utilizados para tareas de clasificación y regresión. Su eficacia en espacios de alta dimensión y su capacidad para manejar datos complejos las hacen particularmente relevantes en aplicaciones médicas, como la detección temprana de enfermedades cerebrovasculares mediante el análisis de imágenes de fondo de ojo. A continuación, se presentan las ventajas y desventajas de aplicar SVM en este contexto:

En cuanto a las **ventajas**, tenemos que como se menciona en FlyRank (s.f.), se destaca:

- ☑ EFICACIA EN ESPACIOS DE ALTA DIMENSIÓN: Las SVM son especialmente

efectivas cuando el número de características supera al número de muestras, una situación común en el análisis de imágenes médicas. Esta capacidad permite manejar datos complejos y de alta dimensionalidad sin comprometer el rendimiento del modelo.

- ☑ CAPACIDAD PARA MODELAR RELACIONES NO LINEALES: Mediante el uso de funciones kernel, las SVM pueden transformar datos no linealmente separables en un espacio de mayor dimensión donde se pueden aplicar técnicas lineales, facilitando la clasificación de patrones complejos presentes en las imágenes de fondo de ojo.
- ☑ ROBUSTEZ FRENTE AL SOBREAJUSTE: Al enfocarse en maximizar el margen entre clases y utilizar solo un subconjunto de los datos de entrenamiento (vectores de soporte), las SVM tienden a generalizar bien en datos no vistos, reduciendo el riesgo de sobreajuste.
- ☑ APLICACIONES EXITOSAS EN EL ANÁLISIS DE IMÁGENES MÉDICAS: Las SVM se han utilizado con éxito en la clasificación de imágenes médicas, incluyendo la detección de retinopatía diabética y otras enfermedades oculares, demostrando su eficacia en la identificación de patrones relevantes en imágenes de fondo de ojo. Por mencionar algunos:
    - *Diagnóstico de melanoma cutáneo*: Un estudio empleó SVM para clasificar imágenes digitales de lesiones cutáneas, diferenciando entre melanoma maligno y nevus displásico. El modelo alcanzó una precisión del 94.1%, superando a otros métodos como el análisis discriminante y las redes neuronales. (Maglogiannis, I. & Zafiropoulos, E., 2004)
    - *Clasificación de tumores cerebrales en imágenes de resonancia magnética (MRI):* Investigaciones han utilizado SVM para distinguir entre tumores cerebrales benignos y malignos en imágenes MRI, logrando una clasificación precisa mediante técnicas de preprocesamiento y extracción de características. (Alrais, R. & Elfadil, N., 2020)
    - (Sharifrazi, D. et al, 2021)

Pero desafortunadamente, también se tienen **desventajas** al aplicarlo (Orlando, 2017):

- ☒ ALTO COSTO COMPUTACIONAL: El entrenamiento de SVM puede ser computacionalmente intensivo, especialmente con grandes conjuntos de datos, debido al cálculo de las funciones kernel y la optimización del hiperplano.
- ☒ SENSIBILIDAD A LA ELECCIÓN DEL KERNEL Y PARÁMETROS: El rendimiento de las SVM depende en gran medida de la selección adecuada de la función kernel y de los parámetros de regularización, lo que puede requerir una extensa validación cruzada.
- ☒ LIMITACIONES EN LA ESTIMACIÓN DE PROBABILIDADES: A diferencia de otros modelos como la regresión logística, las SVM no proporcionan directamente probabilidades de pertenencia a una clase, lo que puede ser una desventaja en aplicaciones clínicas donde se requiere una medida de certeza.
- ☒ MENOR INTERPRETABILIDAD: Aunque las SVM son modelos potentes, su naturaleza matemática y el uso de funciones kernel pueden dificultar la interpretación de los resultados por parte de profesionales clínicos.

Ante esto, podemos decir que las Máquinas de Soporte Vectorial ofrecen una combinación de precisión y robustez que las hace adecuadas para el análisis de imágenes de fondo de ojo en la detección temprana de enfermedades cerebrovasculares. Sin embargo, su aplicación requiere consideraciones cuidadosas respecto al costo computacional, la selección de parámetros y la interpretabilidad de los resultados. En entornos clínicos, es fundamental equilibrar la precisión del modelo con la necesidad de interpretaciones claras y comprensibles para los profesionales de la salud.

Por otro lado, las *Redes Neuronales*, especialmente las redes neuronales convolucionales (CNN), han transformado el análisis de imágenes médicas gracias a su capacidad para aprender representaciones complejas y realizar

tareas de clasificación, segmentación y detección con alta precisión. En el contexto de la detección temprana de enfermedades cerebrovasculares mediante el análisis de imágenes de fondo de ojo, las RNA ofrecen ventajas significativas, aunque también presentan desafíos importantes.

En cuanto a las **ventajas**, tenemos que como se menciona en aws (s.f.), se destaca:

- ☑ CAPACIDAD PARA APRENDER CARACTERÍSTICAS COMPLEJAS: Las RNA pueden identificar patrones sutiles y complejos en las imágenes médicas que podrían pasar desapercibidos para los métodos tradicionales, mejorando así la precisión diagnóstica.
- ☑ AUTOMATIZACIÓN DEL ANÁLISIS DE IMÁGENES: Permiten automatizar tareas como la segmentación de estructuras anatómicas y la detección de anomalías, lo que agiliza el flujo de trabajo clínico y reduce la carga de trabajo de los profesionales de la salud.
- ☑ ADAPTABILIDAD A DIFERENTES TIPOS DE DATOS: Las RNA son versátiles y pueden adaptarse a diversos tipos de datos médicos, incluyendo imágenes de resonancia magnética, tomografías computarizadas y fotografías de fondo de ojo, facilitando su integración en múltiples aplicaciones clínicas.
- ☑ MEJORA CONTINUA MEDIANTE APRENDIZAJE PROFUNDO: A través del aprendizaje profundo, las RNA pueden mejorar su rendimiento con el tiempo al exponerse a más datos, lo que permite una evolución constante en la precisión y eficacia del modelo.

Pero desafortunadamente, también se tienen **desventajas** al aplicarlo (Agostinelli, s.f.):

- ☒ REQUIEREN GRANDES VOLÚMENES DE DATOS ETIQUETADOS: El entrenamiento efectivo de las RNA necesita conjuntos de datos extensos y bien etiquetados, lo que puede ser un desafío en el ámbito médico debido a la disponibilidad limitada de datos y a las consideraciones éticas.
- ☒ ALTA DEMANDA COMPUTACIONAL: El entrenamiento y la implementación de RNA, especialmente las de arquitectura profunda, requieren recursos

computacionales significativos, lo que puede limitar su uso en entornos con infraestructura tecnológica limitada.

- ☒ DIFICULTAD PARA INTERPRETAR LOS RESULTADOS: Las RNA a menudo funcionan como "cajas negras", lo que significa que es difícil entender cómo llegan a una determinada conclusión, lo que puede ser problemático en contextos clínicos donde la interpretabilidad es crucial.
- ☒ RIESGO DE SOBREAJUSTE: Sin una adecuada regularización y validación, las RNA pueden sobreajustarse a los datos de entrenamiento, lo que reduce su capacidad para generalizar a nuevos datos y puede afectar negativamente su rendimiento en la práctica clínica.

En general, podemos ver que las redes neuronales ofrecen un potencial significativo para mejorar la detección temprana de enfermedades cerebrovasculares a través del análisis de imágenes de fondo de ojo, gracias a su capacidad para aprender y automatizar tareas complejas. Sin embargo, su implementación efectiva requiere abordar desafíos relacionados con la necesidad de grandes volúmenes de datos, la demanda computacional y la interpretabilidad de los modelos. Una estrategia cuidadosa que incluya la recolección de datos adecuados, la inversión en infraestructura tecnológica y el desarrollo de métodos para interpretar los resultados de las RNA será esencial para aprovechar al máximo sus beneficios en aplicaciones clínicas.

En cuanto a *k-Vecinos Más Cercanos (k-NN)*, podemos decir que “es una técnica de aprendizaje supervisado ampliamente utilizada en clasificación y regresión. Su simplicidad y eficacia lo hacen atractivo para diversas aplicaciones,” incluyendo el análisis de imágenes médicas para la detección de accidentes cerebrovasculares.

En cuanto a las **ventajas**, tenemos que como se menciona en IBM (s.f.-a), se destaca:

- ☑ SIMPLICIDAD Y FACILIDAD DE IMPLEMENTACIÓN: k-NN es intuitivo y fácil de implementar, lo que lo convierte en una opción accesible para quienes se inician en el aprendizaje automático.

- ☑ NO REQUIERE SUPOSICIONES SOBRE LA DISTRIBUCIÓN DE LOS DATOS: Al ser un algoritmo no paramétrico, k-NN no asume una distribución específica de los datos, lo que le permite adaptarse a diversas estructuras de datos.
- ☑ ADAPTABILIDAD A NUEVOS DATOS: k-NN puede adaptarse fácilmente a nuevos datos sin necesidad de reentrenamiento, ya que utiliza todo el conjunto de datos de entrenamiento para hacer predicciones.
- ☑ APLICACIONES EXITOSAS EN IMÁGENES MÉDICAS: k-NN se ha utilizado con éxito en la clasificación de imágenes médicas, como en el reconocimiento de dígitos escritos a mano y en la clasificación de células en imágenes de aspirados mamarios.
    - *Diagnóstico de cáncer de próstata mediante resonancia magnética (MRI):* Un estudio implementó una técnica basada en la matriz de co-ocurrencia de niveles de gris (GLCM) junto con k-NN para detectar cáncer de próstata en imágenes de MRI. Los resultados mostraron una mejora significativa en la precisión del diagnóstico temprano. (Anand, l. et al, 2024)
    - *Reconocimiento automático de órganos abdominales en imágenes de ultrasonido.* Investigadores desarrollaron un método que combina redes neuronales profundas con k-NN para identificar automáticamente órganos abdominales en imágenes de ultrasonido. El sistema logró una precisión del 96.67% en la clasificación en tiempo real. (Li, K., Xu, Y. & Heng, M, 2021)
    - *Clasificación de patrones de enfermedades pulmonares intersticiales en tomografías computarizadas (CT).* Se utilizó k-NN junto con características topológicas de textura para clasificar patrones morfológicos en imágenes de CT de pacientes con enfermedades pulmonares intersticiales. El enfoque alcanzó una precisión del 97.5% en la identificación de tejidos patológicos. (Huber, M. et al, 2010)

Pero desafortunadamente, también se tienen **desventajas** al aplicarlo (Agostinelli, s.f.):

- ☒ ALTA DEMANDA COMPUTACIONAL: k-NN requiere calcular la distancia entre el punto de prueba y todos los puntos del conjunto de entrenamiento, lo que puede ser costoso en términos de tiempo y memoria, especialmente con conjuntos de datos grandes.
- ☒ SENSIBILIDAD A LA DIMENSIONALIDAD: El rendimiento de k-NN puede degradarse en espacios de alta dimensión debido a la "maldición de la dimensionalidad", donde las distancias entre puntos se vuelven menos significativas.
- ☒ NECESIDAD DE ESCALADO DE CARACTERÍSTICAS: Las características con diferentes escalas pueden influir desproporcionadamente en el cálculo de distancias, por lo que es necesario normalizar o estandarizar los datos antes de aplicar k-NN.
- ☒ SENSIBILIDAD AL RUIDO Y A DATOS ATÍPICOS: k-NN puede verse afectado por datos ruidosos o atípicos, ya que considera los vecinos más cercanos sin evaluar la calidad de los datos.

De esto, podemos decir que k-NN es un algoritmo versátil y fácil de implementar que puede ser útil en la detección de accidentes cerebrovasculares mediante el análisis de imágenes médicas. Sin embargo, su aplicabilidad puede verse limitada por su alta demanda computacional y sensibilidad a la dimensionalidad y al ruido. Es importante considerar estas limitaciones y aplicar técnicas de preprocesamiento adecuadas para mitigar sus desventajas.

Como penúltimo modelo de aprendizaje supervisado tenemos *Random Forest*, que es una técnica de aprendizaje automático basada en conjuntos de árboles de decisión que ha demostrado ser eficaz en diversas aplicaciones médicas, incluyendo la detección y predicción de accidentes cerebrovasculares.

En cuanto a las **ventajas**, tenemos que como se menciona en el artículo de Lavanya, S & Subbulakshmi, P. (2024), se destaca:

- ☑ MANEJO DE DATOS DE ALTA DIMENSIONALIDAD: RF puede procesar conjuntos de datos con un gran número de variables, lo que es común en imágenes médicas y datos clínicos complejos.

- ☑ REDUCCIÓN DEL SOBREAJUSTE: Al combinar múltiples árboles de decisión y utilizar técnicas como el bagging, RF reduce la probabilidad de sobreajuste en comparación con un solo árbol de decisión.
- ☑ CAPACIDAD PARA MANEJAR DATOS FALTANTES Y DESEQUILIBRADOS: RF es robusto frente a datos faltantes y puede manejar conjuntos de datos desequilibrados mediante técnicas de ponderación y muestreo.
- ☑ APLICACIONES EXITOSAS EN LA PREDICCIÓN DE ACCIDENTES CEREBROVASCULARES: Estudios han demostrado que RF puede predecir con alta precisión la incidencia y los resultados de accidentes cerebrovasculares. Por ejemplo, un estudio logró una precisión del 98% en la predicción de accidentes cerebrovasculares utilizando RF.
    - *Predicción temprana de accidentes cerebrovasculares con alta precisión.* Un estudio reciente identificó a Random Forest como el modelo óptimo para la predicción efectiva de accidentes cerebrovasculares, alcanzando una precisión del 98%. Este resultado subraya la capacidad de RF para intervenir de manera temprana y precisa en la detección de eventos cerebrovasculares. (Fernández-Lozano C., et al., 2021)
    - *Predicción de resultados funcionales en pacientes con hemorragia intracerebral primaria.* La técnica de aprendizaje automático utilizando Random Forest predice con precisión los resultados funcionales en pacientes con hemorragia intracerebral primaria a los 1 y 6 meses, lo que ayuda en las decisiones clínicas y la atención al paciente. (Wang, H., et al. 2019)
    - *Predicción de resultados clínicos en pacientes con accidente cerebrovascular isquémico agudo tratados con trombólisis.* Un estudio piloto en el sudeste asiático desarrolló un modelo de aprendizaje automático para predecir el resultado clínico de pacientes con accidente cerebrovascular isquémico agudo después de la trombólisis. El modelo de Random Forest demostró un rendimiento prometedor en la predicción del resultado clínico, medido por la

escala de Rankin modificada (mRS) después de tres meses. (Yunus, R. et al., 2024)

Pero desafortunadamente, también se tienen **desventajas** al aplicarlo (Fernández-Lozano C., et al., 2021) e (Ismail, L., & Materwala, H., 2023):

- ☒ MENOR INTERPRETABILIDAD: Aunque RF proporciona métricas de importancia de variables, la interpretación de sus resultados puede ser más compleja que la de modelos más simples como la regresión logística.
- ☒ REQUIERE MÁS RECURSOS COMPUTACIONALES: Debido a la construcción de múltiples árboles, RF puede ser más exigente en términos de tiempo de entrenamiento y recursos computacionales.
- ☒ SENSIBILIDAD A DATOS ALTAMENTE DESEQUILIBRADOS: En conjuntos de datos con clases muy desequilibradas, RF puede favorecer la clase mayoritaria, lo que afecta la precisión en la detección de eventos raros como ciertos tipos de accidentes cerebrovasculares.
- ☒ POSIBLE DEGRADACIÓN DEL RENDIMIENTO CON VARIABLES ALTAMENTE CORRELACIONADAS: La presencia de variables altamente correlacionadas puede afectar la eficacia de RF, ya que puede seleccionar repetidamente las mismas variables en diferentes árboles.

Dado esto, observamos que el modelo Random Forest representa una opción robusta y efectiva para aplicar en el contexto del análisis de imágenes médicas, específicamente en la detección temprana de enfermedades cerebrovasculares a través del análisis de fondo de ojo. Su capacidad para manejar grandes volúmenes de datos, tolerar valores faltantes, y reducir el sobreajuste lo convierten en un candidato ideal para trabajar con bases de datos clínicas e imágenes de alta dimensionalidad, como las generadas por cámaras retinianas o sistemas OCT (Tomografía de Coherencia Óptica).

Por último, tenemos *Gradient Boosting (XGBoost, LightGBM)*, el cual “es un algoritmo de ensamble basado en árboles de decisión que construye modelos de forma secuencial, corrigiendo los errores del modelo anterior en cada paso. Sus versiones mejoradas como XGBoost (Extreme Gradient Boosting) y LightGBM

(Light Gradient Boosting Machine) han demostrado ser especialmente poderosas en tareas de clasificación médica y análisis de imágenes."

En cuanto a las **ventajas**, se destacan:

- ☑ ALTO RENDIMIENTO PREDICTIVO: Gradient Boosting, especialmente XGBoost y LightGBM, suelen obtener los mejores resultados en competiciones de ciencia de datos y tareas reales de clasificación médica, gracias a su capacidad para capturar patrones complejos. Chen & Guestrin, 2016).
- ☑ MANEJO EFICIENTE DE DATOS NO BALANCEADOS Y VALORES FALTANTES: LightGBM incluye parámetros para manejar clases desequilibradas y valores faltantes, lo que es común en bases de datos médicas (Ke et al., 2017).
- ☑ VELOCIDAD Y EFICIENCIA: XGBoost y LightGBM están optimizados para ejecución rápida y consumo eficiente de memoria, lo que permite entrenar modelos con millones de muestras sin un alto costo computacional.
- ☑ APLICACIONES EXITOSAS EN SALUD Y NEUROCIENCIA:
    - XGBoost se ha usado para predecir resultados de accidentes cerebrovasculares a partir de datos clínicos y biomarcadores (Chen et al., 2024).
    - LightGBM ha sido aplicado en la detección de retinopatía diabética, que también utiliza imágenes de fondo de ojo (Wang et al., 2023).
    - En un estudio reciente, LightGBM superó a Random Forest y SVM en la predicción de mortalidad por ictus isquémico (Zhou et al., 2025).

Pero desafortunadamente, también se tienen **desventajas** al aplicarlo (Lundberg & Lee, 2017):

- ☒ MENOR INTERPRETABILIDAD QUE MODELOS SIMPLES: Aunque existen herramientas como SHAP para interpretar modelos de boosting, no son tan intuitivos como una regresión logística o un árbol de decisión individual.
- ☒ SENSIBILIDAD A HIPERPARÁMETROS: El rendimiento de Gradient Boosting depende fuertemente del ajuste de múltiples parámetros (número de árboles, tasa de aprendizaje, profundidad), lo que puede requerir validación

cruzada intensiva.

- ☒ PROPENSO A SOBREAJUSTE SI NO SE REGULA: Modelos mal ajustados pueden sobreentrenarse fácilmente sobre datos ruidosos o pequeños.
- ☒ REQUIERE MÁS TIEMPO DE ENTRENAMIENTO QUE MODELOS MÁS SIMPLES: Aunque más rápido que otros métodos complejos (como redes neuronales profundas), sigue siendo más lento que Random Forest o regresión logística.

Como podemos ver Gradient Boosting, particularmente en sus versiones XGBoost y LightGBM, ofrece una excelente capacidad predictiva y ha sido probado con éxito en entornos clínicos y de imagenología médica. Para un proyecto de detección temprana de enfermedades cerebrovasculares mediante imágenes de fondo de ojo, es una de las opciones más prometedoras, especialmente si se cuenta con una infraestructura computacional adecuada y se acompaña con técnicas de interpretación de modelos.

Su alto rendimiento justifica su uso, especialmente en etapas avanzadas del pipeline o como parte de un sistema de ensamble.

Personalmente de los modelos vistos, me quedaría con 3 para aplicarlos al proyecto, dado que se requiere alta precisión, capacidad para manejar datos complejos (como imágenes), y cierta explicabilidad, los tres modelos de aprendizaje supervisado más adecuados son:

1. ***GRADIENT BOOSTING (XGBOOST / LIGHTGBM)***, por su rendimiento predictivo superior en múltiples tareas médicas, incluyendo análisis de imágenes y diagnóstico clínico. Además, de que versiones como LightGBM son rápidas y eficientes, ideales para grandes volúmenes de datos. Otra ventaja es que permite manejar datos desequilibrados, comunes en contextos clínicos (desequilibrio entre clases sanas y patológicas).

   Etapas avanzadas del flujo de trabajo (producción), sistemas de apoyo a decisiones clínicas.
2. ***REDES NEURONALES***, dado que son considerados los modelos más eficaces en análisis de imágenes, especialmente convolucionales (CNNs), que son

el estándar en visión por computadora médica. Además, son capaces de detectar patrones visuales complejos que podrían pasar desapercibidos a simple vista o por otros modelos y son fáciles de escalar con grandes bases de datos.

Extracción automática de características desde imágenes crudas, diagnóstico automatizado y sistemas inteligentes de cribado.

3. ***RANDOM FOREST***, porque a pesar de ser un modelo robusto, es fácil de usar y difícil de sobreajustar, lo que permite tener una alta precisión en tareas clínicas cuando se usan variables tabulares (biomarcadores, edad, presión arterial, etc.), resultando muy útil como modelo base o de comparación y también para generar características importantes (feature selection)

   Análisis de datos tabulares clínicos complementarios a las imágenes, prototipado rápido, y en conjuntos de datos con ruido o valores faltantes.

Pasando a los *Modelos de Aprendizaje No Supervisado*, tenemos que estos como se menciona en IBM (s.f.-b) “utiliza datos sin etiquetar. A partir de esos datos, descubre patrones que ayudan a resolver problemas de agrupamiento o asociación. Esto es particularmente útil cuando los expertos no están seguros de las propiedades comunes dentro de un conjunto de datos”. A continuación se abordarán algunos de estos modelos.

Como primer modelo se tiene *Agrupamiento k-means*, el cual es un algoritmo de agrupamiento (clustering) que divide un conjunto de datos en k grupos (clusters) basándose en la distancia euclidiana entre los puntos y sus centroides. Es uno de los métodos más usados en aprendizaje no supervisado debido a su simplicidad y eficiencia.

En cuanto a las **ventajas**, se destacan:

- ☑ SIMPLICIDAD Y RAPIDEZ COMPUTACIONAL: K-Means es muy rápido y escalable a grandes volúmenes de datos, lo cual es útil al trabajar con conjuntos grandes de imágenes médicas sin etiquetas.
- ☑ PERMITE EXPLORACIÓN Y PREPROCESAMIENTO DE DATOS: Es útil para descubrir patrones ocultos en los datos y realizar agrupaciones preliminares, lo que

puede ayudar en la selección de regiones de interés (ROIs) en las imágenes del fondo de ojo.

- ☑ Apoya la detección no supervisada de anomalías: Puede identificar subconjuntos de pacientes con características similares, incluso sin etiquetas previas. Esto puede orientar hipótesis sobre subtipos clínicos o biomarcadores. (Oliveira et al., 2016)
- ☑ Aplicaciones exitosas en imágenes médicas: Se ha utilizado en la segmentación de vasos sanguíneos, regiones afectadas por edema, y para el agrupamiento de patrones de retinopatía.

Pero desafortunadamente, también se tienen **desventajas** al aplicarlo (Jain, 2010):

- ☒ Necesita definir el número de clústeres (k) a priori: Esto puede ser problemático si no se conoce la cantidad óptima de grupos en los datos, y elegir un k incorrecto puede sesgar el análisis.
- ☒ Sensibilidad a la inicialización y outliers: La calidad de los resultados depende de la elección inicial de los centroides y puede verse afectada por datos atípicos.
- ☒ Asume que los clústeres son esféricos y del mismo tamaño: Esta suposición puede ser poco realista en el contexto médico, donde las características visuales pueden variar en forma y distribución.
- ☒ No captura relaciones complejas: K-Means es limitado para datos con estructuras no lineales, como imágenes en bruto, sin reducción de dimensionalidad previa.

De esto podemos ver que K-Means puede ser una herramienta útil en las etapas iniciales del pipeline de análisis, especialmente para exploración de datos no etiquetados, segmentación preliminar de regiones en imágenes retinianas y/o agrupamiento de pacientes o patrones visuales similares.

Sin embargo, no debe utilizarse como modelo final de predicción diagnóstica, sino como complemento a modelos supervisados, o para asistir en la generación de etiquetas (semi-supervisado) y reducción de dimensionalidad.

Por su parte para *Clustering Jerárquico*, tenemos que es un método de agrupamiento que construye una jerarquía de clústeres. Puede hacerse de dos formas, por un lado, Aglomerativo (bottom-up), es decir, cada observación inicia como un clúster individual, y se van agrupando. O de manera, divisiva (top-down), es decir, parte de un clúster grande que se divide recursivamente.
El resultado puede visualizarse con un dendrograma, que muestra cómo se agrupan los datos.

En cuanto a las **ventajas**, se destacan (Müller, 2011):

- ☑ NO REQUIERE ESPECIFICAR EL NÚMERO DE CLÚSTERES (K) A PRIORI: A diferencia de K-Means, permite observar la estructura de los datos y decidir luego el número óptimo de grupos.
- ☑ PRODUCE UNA REPRESENTACIÓN VISUAL INTUITIVA: El dendrograma permite interpretar de manera visual las relaciones entre instancias, útil en contextos médicos donde la trazabilidad de decisiones es relevante (Kassambara, 2017).
- ☑ DETECCIÓN DE PATRONES JERÁRQUICOS COMPLEJOS: Puede capturar subgrupos anidados, lo cual es útil si hay variaciones sutiles entre tipos de pacientes o lesiones retinianas.
- ☑ APLICACIONES EN ANÁLISIS BIOMÉDICO Y DE IMÁGENES: Ha sido utilizado para:
  - Identificación de subtipos de enfermedades neurológicas.
  - Agrupamiento de características extraídas de imágenes de fondo de ojo en análisis de retinopatías o glaucoma.

Pero desafortunadamente, también se tienen **desventajas** al aplicarlo (Murtagh, F., & Contreras, 2012):

- ☒ ESCALABILIDAD LIMITADA: Es computacionalmente costoso para grandes conjuntos de datos (especialmente aglomerativo con distancias completas), lo que puede ser un problema con miles de imágenes de fondo de ojo.
- ☒ SENSIBILIDAD AL TIPO DE MÉTRICA DE DISTANCIA: Los resultados pueden variar significativamente dependiendo de si se usa distancia euclidiana, Manhattan, etc.

- ☒ DIFÍCIL DE ACTUALIZAR: Una vez construido el dendrograma, no es fácil integrar nuevos datos sin rehacer todo el modelo.
- ☒ MENOR APLICABILIDAD DIRECTA AL ANÁLISIS DE IMÁGENES CRUDAS: Generalmente requiere una reducción de dimensionalidad previa (como PCA o autoencoders) para poder trabajar con imágenes, ya que no escala bien con datos de alta dimensión.

En general, vemos que el Clustering Jerárquico es una herramienta útil para la exploración de relaciones estructurales complejas en los datos, y puede ser especialmente valiosa para interpretar agrupaciones de pacientes o patrones clínicos si cuentas con características tabulares o embeddings extraídos de imágenes.
A pesar de esto, no es recomendable usarlo sobre las imágenes en crudo, pero sí puede ser muy potente al trabajar sobre representaciones reducidas o características clínicas derivadas.

Como siguiente modelo tenemos, *DBSCAN (Density-Based Spatial Clustering of Applications with Noise)*, el cual es un algoritmo de agrupamiento que forma clústeres en función de regiones densamente pobladas de puntos, detectando automáticamente regiones de ruido y formas arbitrarias. No requiere que especifiques el número de clústeres de antemano, lo que lo hace valioso en contextos donde no se conoce la estructura de los datos.

En cuanto a las **ventajas**, se destacan:

- ☑ DETECTA CLÚSTERES DE FORMAS ARBITRARIAS: A diferencia de K-Means, permite observar la estructura de los datos y decidir luego el número óptimo de grupos.
- ☑ IDENTIFICACIÓN AUTOMÁTICA DE RUIDO Y OUTLIERS: Tiene la capacidad de marcar automáticamente puntos que no pertenecen a ningún grupo como “ruido”, lo cual es útil en imágenes médicas para detectar anomalías o lesiones atípicas (Sander et al., 1998).
- ☑ NO REQUIERE ESPECIFICAR EL NÚMERO DE CLÚSTERES: Solo requiere dos

parámetros (ε y minPts), y el número de clústeres se determina con base en la densidad de los datos, lo cual favorece exploraciones no supervisadas.

- ☑ APLICACIONES EXITOSAS EN IMÁGENES MÉDICAS: DBSCAN ha sido usado en segmentación de estructuras retinianas, clasificación de microaneurismas, y análisis de estructuras cerebrales en resonancia magnética (Lin et al., 2017).

Pero desafortunadamente, también se tienen **desventajas** al aplicarlo:

- ☒ SENSIBILIDAD A LOS PARÁMETROS (ε Y minPts): Es computacionalmente costoso para grandes conjuntos de datos (especialmente aglomerativo con distancias completas), lo que puede ser un problema con miles de imágenes de fondo de ojo.
- ☒ PROBLEMAS DE ESCALABILIDAD: DBSCAN tiene una complejidad de $O(n^2)$ en implementaciones básicas, lo que puede ser ineficiente para bases de datos muy grandes (Guo, et al., 2017).
- ☒ DIFÍCIL APLICACIÓN DIRECTA SOBRE IMÁGENES CRUDAS: Necesita que las imágenes sean primero representadas como vectores de características (por ejemplo, mediante extracción con CNNs o PCA), ya que no trabaja bien en espacios de alta dimensionalidad.
- ☒ NO ES DETERMINÍSTICO CON DATOS COMPLEJOS O RUIDOSOS: Pequeñas variaciones pueden llevar a resultados distintos si los datos son densamente poblados de forma heterogénea.

En general podemos ver que DBSCAN es una opción poderosa para análisis exploratorios no supervisados, especialmente si se trata de segmentar regiones anómalas en imágenes de fondo de ojo, detectar grupos de pacientes con patrones retinianos inusuales y/o filtrar ruido o imágenes irrelevantes en etapas de preprocesamiento.

Sin embargo, no es adecuado como modelo principal de diagnóstico automático, pero sí como complemento en la etapa de detección preliminar, extracción de clústeres o limpieza de datos.

Cambiando de modelo, tenemos *Análisis de Componentes Principales (PCA)* el cual es una técnica estadística que transforma un conjunto de variables posiblemente correlacionadas en un nuevo conjunto de variables linealmente independientes llamadas componentes principales. Su objetivo es reducir la dimensionalidad del conjunto de datos maximizando la varianza explicada, conservando así la mayor cantidad posible de información.

En cuanto a las **ventajas**, se destacan:

- ☑ REDUCCIÓN EFECTIVA DE DIMENSIONALIDAD EN IMÁGENES: Las imágenes médicas (como los fondos de ojo) suelen tener miles de características por píxel. PCA permite condensar esa información conservando las características más relevantes, lo cual es útil antes de aplicar modelos supervisados o clustering.
- ☑ MEJORA EL RENDIMIENTO DE MODELOS POSTERIORES: Al eliminar redundancias y ruido, mejora la eficiencia de algoritmos como SVM, k-NN o DBSCAN, al trabajar en un espacio más compacto y menos ruidoso (Shlens, 2014).
- ☑ VISUALIZACIÓN DE DATOS COMPLEJOS: Con dos o tres componentes principales, PCA facilita visualizar estructuras o agrupaciones naturales en los datos, útil en análisis exploratorio (Jolliffe & Cadima, 2016).
- ☑ APLICACIONES EXITOSAS EN ANÁLISIS MÉDICO: Se ha aplicado para:
    - Clasificación de retinopatías.
    - Identificación de características discriminantes en resonancias magnéticas cerebrales.
    - Segmentación automática de vasos y lesiones retinales (Karunanayake & Kodikara, 2015).

Pero desafortunadamente, también se tienen **desventajas** al aplicarlo (Lever et al., 2017):

- ☒ PÉRDIDA DE INTERPRETABILIDAD: Los componentes principales son combinaciones lineales de características originales, lo que dificulta su interpretación clínica directa.
- ☒ LINEALIDAD: PCA asume que la estructura de los datos es lineal. En imágenes complejas, relaciones no lineales podrían quedar fuera (ideal

usar alternativas como t-SNE o autoencoders si este es el caso).

- ☒ ESCALABILIDAD Y NORMALIZACIÓN: Requiere estandarización previa y puede ser costoso computacionalmente para conjuntos muy grandes si no se usa una versión optimizada (como Incremental PCA o Truncated SVD).
- ☒ NO REALIZA AGRUPAMIENTO NI PREDICCIÓN: PCA no es un modelo de clustering o clasificación per se; requiere ser combinado con otros modelos para tareas predictivas.

Ante esto podemos decir que PCA no es un modelo predictivo en sí, sino una herramienta de preprocesamiento esencial. En este caso el aplicarlo en el proyecto puede ser altamente útil para reducir la dimensionalidad de imágenes de fondo de ojo antes de aplicar modelos de detección (por ejemplo, SVM o k-NN), eliminar ruido o información redundante y/o facilitar la visualización y comprensión de los patrones latentes en los datos.
También es recomendable como paso previo a cualquier modelo supervisado o de clustering si se trabaja con imágenes transformadas en vectores de características.

Como siguiente modelo tenemos Autoencoders, los cuales son un tipo de red neuronal artificial diseñada para aprender una representación comprimida (codificación) de los datos. Consisten en dos partes:

I. Codificador (encoder): reduce la dimensionalidad aprendiendo las características más relevantes.
II. Decodificador (decoder): intenta reconstruir los datos originales a partir de la representación comprimida.

Estos se entrenan en modo no supervisado, minimizando el error de reconstrucción.

En cuanto a sus **ventajas**, se destacan:

- ☑ REDUCCIÓN DE DIMENSIONALIDAD NO LINEAL: A diferencia de PCA, los autoencoders pueden capturar relaciones no lineales complejas en los datos, lo cual es clave para trabajar con imágenes médicas que contienen

patrones estructurales sutiles y variados (Goodfellow et al., 2016).

- ☑ Extracción automática de características (feature learning): Aprenden representaciones abstractas útiles que pueden ser utilizadas por modelos supervisados o de clustering, lo cual automatiza la ingeniería de características, mejorando la eficiencia del pipeline de análisis (Hinton & Salakhutdinov, 2006).
- ☑ Detección de anomalías: Autoencoders pueden identificar imágenes anómalas (como posibles signos de enfermedad) al presentar altos errores de reconstrucción, algo útil en detección temprana de patologías (Chen & Konukoglu, 2018).
- ☑ Aplicaciones exitosas en imágenes de fondo de ojo: Han sido utilizados en:
  - Segmentación de vasos retinianos.
  - Detección de retinopatía diabética (Ortiz-Feregrino et al., 2022),
  - Reducción de ruido en imágenes OCT o fondo de ojo.

Pero desafortunadamente, también se tienen **desventajas** al aplicarlo (Lever et al., 2017):

- ☒ ENTRENAMIENTO COSTOSO Y SENSIBLE: Requiere gran cantidad de datos y poder computacional (especialmente en versiones profundas), además de una cuidadosa selección de arquitectura y regularización para evitar el sobreajuste (LeCun et al., 2015).
- ☒ INTERPRETABILIDAD LIMITADA: Las representaciones aprendidas por el codificador son difíciles de interpretar clínicamente, lo cual puede limitar su aceptación en entornos médicos sensibles (Samek et al., 2017).
- ☒ NO ES UN MODELO DE PREDICCIÓN DIRECTA: Al igual que PCA, los autoencoders son una técnica de preprocesamiento o detección de patrones latentes, no sirven directamente como clasificadores (aunque sus embeddings sí pueden alimentar modelos como SVM o XGBoost).
- ☒ REQUIEREN DATOS LIMPIOS Y BIEN ETIQUETADOS PARA MEJORES RESULTADOS: Aunque no supervisados, el entrenamiento efectivo depende de buenos datos sin ruido sistemático o sesgo, algo que puede ser un reto en bases de imágenes médicas (Litjens et al., 2017).

El penúltimo modelo es *Mapas Autoorganizados (SOM)*, los cuales son una clase especial de redes neuronales artificiales desarrollada por Teuvo Kohonen. Su objetivo principal es proyectar datos de alta dimensión a un espacio de menor dimensión (usualmente bidimensional), preservando la topología del espacio de características original.

En cuanto a sus **ventajas**, se destacan:

- ☑ VISUALIZACIÓN INTUITIVA DE DATOS COMPLEJOS: SOM permite representar datos multivariados de alta dimensión en mapas bidimensionales fácilmente interpretables, lo que ayuda a descubrir agrupamientos y relaciones ocultas (Kohonen, 2001).
- ☑ PRESERVACIÓN DE LA TOPOLOGÍA: A diferencia de métodos como k-means, SOM mantiene la proximidad entre vectores similares en el espacio de entrada, lo que es útil para visualizar patrones de enfermedad o grupos de pacientes con características similares (Vesanto & Alhoniemi, 2000).
- ☑ NO REQUIERE ETIQUETADO: Funciona de forma completamente no supervisada, ideal cuando se tiene una gran cantidad de imágenes no anotadas —como ocurre frecuentemente en bases de datos médicas (Estévez et al., 2009).
- ☑ APLICACIONES EXITOSAS EN MEDICINA Y OFTALMOLOGÍA: SOM ha sido utilizado en:
    - Agrupamiento de pacientes con enfermedades neurodegenerativas.
    - Análisis de patrones en electroencefalogramas y resonancias cerebrales (Ritter et al., 1992).

Pero desafortunadamente, también se tienen **desventajas** al aplicarlo:

- ☒ SENSIBILIDAD A LA INICIALIZACIÓN Y PARÁMETROS: La calidad de los mapas depende del tamaño del mapa, tasa de aprendizaje y número de épocas. Un mal ajuste puede llevar a resultados poco estables o irrelevantes (Bianchesi et al., 2022).
- ☒ DIFICULTAD EN INTERPRETAR LOS CLÚSTERES CLÍNICAMENTE: Aunque el mapa

es visual, su interpretación médica requiere correlación con variables clínicas específicas, lo que puede no ser inmediato (Carpenter & Grossberg, 1987).

- ☒ NO REALIZA CLASIFICACIÓN DIRECTA: SOM solo agrupa, no clasifica. Si se desea predecir la presencia o no de enfermedad, se requiere complementar con un clasificador supervisado (e.g., SVM o XGBoost sobre las salidas de SOM).

En general podemos decir que SOM es una herramienta útil para la exploración y visualización de datos complejos de fondo de ojo, especialmente cuando se desea entender la estructura latente de los datos, se tiene poca información etiquetada o se busca agrupar pacientes o imágenes por similitud sin intervención humana.

Sin embargo, su uso como herramienta principal de predicción es limitado. Es más adecuado como paso exploratorio o como complemento previo a un modelo supervisado.

Y como último modelo, tenemos ***Reducción de Dimensionalidad con t-SNE***, es un método no supervisado de reducción de dimensionalidad no lineal desarrollado por van der Maaten y Hinton (2008). Su objetivo es preservar la estructura local de los datos al proyectarlos en un espacio de 2 o 3 dimensiones, permitiendo visualizar agrupamientos latentes o anomalías en datos complejos.

En cuanto a sus **ventajas**, se destacan:

- ☑ VISUALIZACIÓN CLARA DE PATRONES Y AGRUPACIONES COMPLEJAS: Es una de las técnicas más eficaces para visualizar clústeres naturales en datos de alta dimensión, como las características extraídas de imágenes retinianas, revelando estructuras útiles para análisis clínicos exploratorios (van der Maaten & Hinton, 2008).
- ☑ Captura relaciones no lineales: A diferencia de métodos lineales como PCA, t-SNE preserva relaciones complejas y distingue mejor entre clases no linealmente separables, lo cual puede ser esencial al buscar diferencias sutiles en imágenes normales y con signos de enfermedad (Wattenberg et

al., 2016).

- ☑ Complemento útil a modelos supervisados: Si bien no clasifica, t-SNE puede usarse para validar la separación entre clases tras la extracción de características con CNNs, autoencoders u otros métodos, ayudando a evaluar visualmente la calidad del embedding.
- ☑ Aplicaciones exitosas en imágenes médicas: Se ha aplicado exitosamente en:
    - Visualización de embeddings de imágenes de retina y OCT (Wang et al., 2019),
    - Análisis de datos genómicos y expresión de proteínas (Becht et al., 2018),
    - Clasificación visual de tumores o enfermedades degenerativas.

Pero desafortunadamente, también se tienen **desventajas** al aplicarlo:

- ☒ COMPUTACIONALMENTE COSTOSO: Tiene un alto costo computacional, especialmente con grandes conjuntos de datos (>10,000 muestras), lo que puede limitar su uso sin reducción previa o submuestreo (van der Maaten, 2008).
- ☒ NO ES INTERPRETABLE CLÍNICAMENTE: Aunque útil para visualización, los ejes generados por t-SNE no tienen significado explícito, lo que puede dificultar la interpretación médica directa.
- ☒ NO ES REPRODUCIBLE: t-SNE es sensible a la inicialización y a los hiperparámetros (como la perplejidad), y puede producir diferentes resultados en cada ejecución, afectando la robustez del análisis si no se controla (Wattenberg et al., 2016).
- ☒ NO APTO PARA CLASIFICACIÓN DIRECTA: t-SNE no es un clasificador ni un modelo de predicción. No se puede usar directamente para asignar etiquetas ni tomar decisiones clínicas; su valor reside en la exploración visual de los datos.

En general podemos decir que t-SNE es una excelente herramienta para el análisis exploratorio y visualización de datos de alta dimensión en proyectos médicos. Para el proyecto, puede ser muy útil para visualizar separaciones entre

grupos de pacientes (sanos vs. enfermos), explorar la calidad de embeddings generados por autoencoders o CNNs y/o etectar posibles anomalías o agrupamientos inesperados en las imágenes de fondo de ojo.
Es importante resaltar que no debe utilizarse como modelo principal de predicción, pero sí como un complemento analítico potente.

Después de haber analizado los modelos, colocó top 3 los modelos que personalmente considero más adecuados:

1. ***AUTOENCODERS***, son ideales para extraer representaciones comprimidas y relevantes de imágenes médicas, facilitando la detección de patrones que podrían indicar condiciones cerebrovasculares incipientes, incluso en ausencia de etiquetas. En este caso se utilizarían como paso previo de extracción de características, seguido de un modelo como SVM, Random Forest o XGBoost para clasificación.
2. ***SELF-ORGANIZING MAPS (SOM)***, como son potentes para visualizar datos complejos y explorar relaciones ocultas en las características de imágenes oftálmicas. Son especialmente útiles cuando se tiene un conjunto de datos no etiquetado o semi-etiquetado. Sería útil en caso de que existan patrones visuales consistentes entre grupos, por ejemplo imágenes normales vs anómalas, esto antes de clasificar.
3. ***T-SNE (T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING)***, ya que es excelente para la visualización exploratoria de datos de imágenes de fondo de ojo, permitiendo identificar si hay separaciones naturales entre grupos, como pacientes sanos y aquellos con riesgo cerebrovascular. En el proyecto ayudaría a visualizar los embeddings extraídos por autoencoders o CNNs, validando si existe diferenciación visual entre clases o pacientes.

PROPUESTA DE APLICACIÓN DE MODELOS

1) *Preprocesamiento y Extracción de Características*, en este paso se busca reducir el ruido, extraer las características relevantes de las imágenes, y facilitar el aprendizaje posterior. Por lo que se optaría por utilizar un modelo

Autoencoders (no supervisado), con la finalidad de codificar imágenes retinianas en vectores de características representativos. Ideal para grandes volúmenes de datos no etiquetados.

Como posible complemento se consideraría utilizar PCA para reducción de dimensionalidad lineal posterior al uso de autoencoders o Mapas Autoorganizados (SOM) para una primera exploración de agrupamientos.

2) *Exploración y Análisis de Patrones Latentes*, a esta altura se busca descubrir si existen agrupamientos naturales (ejemplo sanos vs. con riesgo) y validar la estructura de los datos. Por lo que se utilizaría modelos como:
    a. t-SNE (no supervisado), con la finalidad de visualizar en 2D o 3D las representaciones comprimidas y detectar grupos visuales de riesgo.
    b. Clustering jerárquico o DBSCAN, para agrupar sin necesidad de etiquetas y observar la distribución de patrones latentes.

3) *Clasificación Supervisada*, donde se buscaría predecir si una imagen de fondo de ojo indica riesgo de accidente cerebrovascular (etiquetas sí/no, o en niveles). Donde se consideraría aplicar ya sea uno o varios de los siguientes modelos:
    a. Random Forest, ya que es robusto, interpretable, tolerante al ruido y con buen desempeño en datos médicos.
    b. XGBoost / LightGBM, puesto que es altamente preciso y eficiente; ideal para conjuntos de datos con muchas características extraídas.
    c. SVM, dado que tiene buen rendimiento en espacios de alta dimensión; excelente para distinguir entre clases separables.
    d. Como una alternativa para clasificación más compleja, se consideraría aplicar redes neuronales profundas (CNNs o MLP), especialmente si se entrena directamente sobre imágenes o embeddings generados por autoencoders.

4) *Interpretabilidad y Validación Clínica*, esto con el objetivo de facilitar la validación médica de los resultados y ayudar en la toma de decisiones clínicas. Para este punto se considera aplicar Random Forest + SHAP o LIME ya que permite obtener explicaciones locales de las predicciones, ayudando a los médicos a entender qué características influyeron más. Otra opción sería aplicar Grad-CAM (si se usan CNNs) para visualizar qué regiones de la imagen fueron más relevantes en la predicción.

ESTRUCTURA DEL TRABAJO EN PYTHON

Se trabajaría mediante módulos .py, los cuales se utilizarían con el objetivo de mantener un orden en los procesos. El primero de ellos sería el *módulo de preprocesamiento* en el que se tratarían las imágenes, en otro módulo se generarían (*model_training*) los modelos mencionados en el punto anterior, en un tercer módulo (*evaluation*) se implementarían los modelos, así como el cálculo de métricas, y por último en un último módulo (*mlops_pipeline*) se implementarían las instrucciones creadas en los otros módulos.

CONCLUSIÓN

La detección temprana de enfermedades cerebrovasculares a través del análisis de imágenes de fondo de ojo representa un desafío técnico y clínico de gran complejidad, que requiere no solo un conocimiento profundo de las técnicas de análisis de datos, sino también una selección estratégica de modelos de aprendizaje automático acordes a cada etapa del proceso.

El análisis detallado de los modelos supervisados y no supervisados ha demostrado que no existe una solución única o universalmente superior. En cambio, la eficacia del sistema depende de un ensamblaje inteligente de modelos, donde cada uno cumple una función específica: desde la reducción de dimensionalidad y extracción de características, hasta la clasificación final y la interpretación clínica de los resultados.

Los modelos no supervisados, como autoencoders, t-SNE y SOM, son herramientas clave en las fases iniciales del pipeline. Permiten comprender la estructura latente de los datos, visualizar agrupamientos naturales y generar representaciones comprimidas que alimenten con eficiencia a los modelos supervisados. Esta capacidad de "preparar el terreno" es esencial, especialmente cuando se dispone de grandes volúmenes de imágenes no etiquetadas.

Por su parte, los modelos supervisados, como Random Forest, XGBoost y SVM, son los más adecuados para la clasificación precisa del riesgo cerebrovascular, dada su capacidad para manejar datos complejos, no lineales y con alta dimensionalidad. Además, su potencial para ser interpretados mediante herramientas como SHAP o Grad-CAM les confiere una ventaja adicional en el entorno clínico, donde la transparencia y la explicabilidad son fundamentales.

En conjunto, este análisis refuerza la idea de que la elección de modelos no debe hacerse de forma aislada, sino como parte de una arquitectura integral alineada con los objetivos clínicos, la naturaleza de los datos y la necesidad de interpretabilidad. Esta estrategia no solo mejora la precisión del sistema, sino que también fortalece su aceptabilidad y utilidad en contextos reales de atención médica.

REFERENCIAS

- Agostinelli, G., Homagi, S. & Alouani, D. (s.f.). ¿Cuáles son las ventajas y desventajas de usar redes neuronales convolucionales para el reconocimiento de imágenes? Recuperado el 15 de mayo de 2025, de https://es.linkedin.com/advice/0/what-advantages-disadvantages-using-convolutional?lang=es
- Alrais, R. & Elfadil, N., (2020, junio). Support Vector Machine (SVM) for Medical Image Classification of Tumorous. International Journal of Computer Science and Mobile Computing. Vol. 9, Issue. 6. Recuperado de https://ijcsmc.com/docs/papers/June2020/V9I6202017.pdf

- Anand, L. et al. (2024, enero 9). Diagnosis of Prostate Cancer Using GLCM Enabled KNN Technique by Analyzing MRI Images. *National Library of Medicine.* Recuperado de https://pubmed.ncbi.nlm.nih.gov/36733405/
- aws. (s.f.). *¿Qué es una red neuronal?* Recuperado el 15 de mayo de 2025, de https://aws.amazon.com/es/what-is/neural-network/
- Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W., Ng, L. G., ... & Newell, E. W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology, 37(1), 38–44*. Recuperado de https://doi.org/10.1038/nbt.4314
- Bianchesi, N., Da Matta, C., Streitenberger, S., Romão, E.., Balestrassi, P., & Costa, A. (2022, septiembre 09). A nonlinear time-series prediction methodology based on neural networks and tracking signals. *ABEPRO.* Recupeado de https://www.redalyc.org/journal/3967/396769689030/html/
- Carpenter, G. A., & Grossberg, S. (1987). ART 2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics, 26(23), 4919–4930*. Recuperado de https://doi.org/10.1364/AO.26.004919
- Chen, T., & Guestrin, C. (2016, Agosto 13). XGBoost: A scalable tree boosting system. *ACM Digital Library*. Recuperado de https://doi.org/10.1145/2939672.2939785
- Chen, X. & Konukoglu, E. (2018, junio 13). Unsupervised Detection of Lesions in Brain MRI using constrained adversarial auto-encoders. *arXiv.* Recuperado de https://arxiv.org/abs/1806.04972
- Chen, H., Wang, H., Yang, C., & Zhao, L. (2024, septiembre 19). Machine learning-based stroke prediction using extreme gradient boosting. *BMC Medical Informatics and Decision Making*, 22, 104. Recuperado de https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-024-02331-1
- FasterCapital. (2025, marzo 29). *Arboles de decision de decodificacion para modelado predictivo preciso.* Recuperado de

https://fastercapital.com/es/contenido/Arboles-de-decision--arboles-de-decision-de-decodificacion-para-modelado-predictivo-preciso.html?utm_source=chatgpt.com

- Fernández-Lozano, C., Hervella, P., Mato-Abad, V., Rodríguez-Yáñez, M., Suárez-Garaboa, S., López-Dequidt, I., Estany-Gestal, A., Sobrino, T., Campos, F., Castillo, J., Rodríguez-Yáñez, S., & Iglesias-Rey, R. (2021, mayo 12). Random Forest-Based Prediction of Stroke Outcome. *Scientific Reports*, 11(1), 89434. Recuperado de https://doi.org/10.1038/s41598-021-89434-7
- FlyRank. (s.f.). *¿Cuáles son las ventajas y desventajas de las máquinas de soporte vectorial?* Recuperado el 15 de mayo de 2025, de https://www.flyrank.com/es/blogs/perspectivas-de-ia/what-are-the-advantages-and-disadvantages-of-support-vector-machines
- Foqum. (s.f.). *¿Qué significa KNN?* Recuperado el 15 de mayo de 2025, de https://foqum.io/blog/termino/knn/
- GeeksforGeeks. (2023). *Advantages and Disadvantages of Logistic Regression.* https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning. MIT Press.* Recuperado de https://www.deeplearningbook.org/
- Godwa, A. & Shorck, J. (s.f.). *What are the advantages and disadvantages of using linear regression for data analysis?* Recuperado el 15 de mayo de 2025, de https://www.linkedin.com/advice/3/what-advantages-disadvantages-using-linear-1c?lang=e
- Guo, Y., et al. (2017, septiembre 06). An Efficient Image Segmentation Algorithm Using Neutrosophic Graph Cut. *MPDI.* Recuperado de https://www.mdpi.com/2073-8994/9/9/185
- Guyatt, G., Wakter, S., Shannon, H., Cook, D., Jaeschke, R. & Heddle, N. (1995). *Estadísticas básicas para médicos.* Recuperado de

https://smiba.org.ar/curso_medico_especialista/lecturas_2021/e%29.%204%20Correlaci%C3%B3n%20y%20regresi%C3%B3n.pdf

- Hinton, G. E., & Salakhutdinov, R. R. (2006, julio 28). Reducing the dimensionality of data with neural networks. *Science, 313(5786), 504–507.* Recuperado de https://doi.org/10.1126/science.1127647
- Huber, M. et al. (2010, mayo 27). Classification of interstitial lung disease patterns with topological texture features. *arXiv.* Recuperado de https://arxiv.org/abs/1005.5086
- IBM. (s.f.-a). *¿Qué es el algoritmo KNN?* Recuperado el 15 de mayo de 2025, de https://www.ibm.com/mx-es/think/topics/knn
- IBM. (s.f.-b). *¿Qué es el aprendizaje supervisado?* Recuperado el 15 de mayo de 2025, de https://www.ibm.com/mx-es/topics/supervised-learning
- IBM. (s.f.-c). *¿Qué es un árbol de decisión?* Recuperado el 15 de mayo de 2025, de https://www.ibm.com/mx-es/think/topics/decision-trees
- Ismail, L., & Materwala, H. (2023, abril 01). From Conception to Deployment: Intelligent Stroke Prediction Framework using Machine Learning and Performance Evaluation. *arXiv*. Recuperado de https://arxiv.org/abs/2304.00249
- Jain, A. (2010, junio 01). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*. Vol 31. Pp.651-666. Recuperado de https://doi.org/10.1016/j.patrec.2009.09.011
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065).* Recuperado de https://doi.org/10.1098/rsta.2015.0202
- Karunanayake, N. & Kodikara, N. (2015, diciembre). An Improved Method for Automatic Retinal Blood Vessel Vascular Segmentation Using Gabor Filter. *Scientific Research.* Recuperado de https://www.scirp.org/journal/paperinformation?paperid=61762

- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146–3154. Recuperado de https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf
- Kohonen, T. (2001). Self-Organizing Maps (3rd ed.). *Springer*. Recuperado de https://doi.org/10.1007/978-3-642-56927-2
- Lavanya, S & Subbulakshmi, P. (2024, agosto 29). Unveiling the potential of machine learning approaches in predicting the emergence of stroke at its onset: a predicting framework. *Scientific reports*. Recuperado de https://www.nature.com/articles/s41598-024-70354-1
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning. Nature, 521(7553), 436–444*. Recuperado de https://doi.org/10.1038/nature14539
- Lever, J., Krzywinski, M., & Altman, N. (2017, junio 29). Principal component analysis. *Nature Methods, 14(7), 641–642*. Recuperado de https://doi.org/10.1038/nmeth.4346
- Li, K., Xu, Y. & Heng, M. (2021, octubre 09). Automatic Recognition of Abdominal Organs in Ultrasound Images based on Deep Neural Networks and K-Nearest-Neighbor Classification. *arXiv.* Recuperado de https://arxiv.org/abs/2110.04563
- Lin, M., et al. (2017, enero 18). Robust Retinal Blood Vessel Segmentation Based on Reinforcement Local Descriptions. *BioMed Research International.* Recuperado de https://pmc.ncbi.nlm.nih.gov/articles/PMC5286479/
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42, 60–88*. Recuperado de https://doi.org/10.1016/j.media.2017.07.005

- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774. Recuperado de https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
- Maglogiannis, I. & Zafiropoulos, E., (2004). Characterization of digital medical images utilizing support vector machines. *BMC Medical Informatics and Decision Making.* Recuperado de https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-4-4
- Müller, D. (2011, septiembre 12). Modern hierarchical, agglomerative clustering algorithms. *arXiv*. Recuperado de https://arxiv.org/abs/1109.2378
- Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: *An overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), 86–97*. Recuperado de https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.53
- Oliveira, W., et al. (2016, febrero 26). Unsupervised Retinal Vessel Segmentation Using Combined Filters. *PLOSOne*. Recuperado de https://doi.org/10.1371/journal.pone.0149943
- Orlando, J. (2017). *Aprendizaje automático para asistencia al diagnóstico de enfermedades visuales basado en imágenes de fondo de ojo.* Instituto Pladema [Tesis de Doctorado]. Recuperado de https://ignaciorlando.github.io/publication/2017-thesis-retina/2017-thesis-retina.pdf
- Ortiz-Feregrino, R., Tovar-Arriaga, S., Pedraza-Ortega, J., & Takacs, A. (2022, cotubre 21). Retinal Lesion Segmentation Using Transfer Learning with an Encoder-Decoder CNN. *Revista mexicana de ingeniería biomédica, 43(2), 1246*. Recuperado de https://doi.org/10.17488/rmib.43.2.4
- Podrez, A. (2023, septiembre 21). *Introducción a la regresión lineal:*

definición y aplicaciones. Universidad de los Andes. Recuperado de https://programas.uniandes.edu.co/blog/regresion-lineal

- RadioCare. (2013, marzo 14). *Ventajas de contar con el apoyo de la Inteligencia Artificial en el área de imagenología.* Recuperado de https://radiocare.mx/blog/ventajas-de-contar-con-el-apoyo-de-la-inteligencia-artificial-en-el-area-de-imagenologia/
- Ritter, H., Martinetz, T., & Schulten, K. (1992). Neural Computation and Self-Organizing Maps: An Introduction. *Addison-Wesley*. Recuperado de https://www.ks.uiuc.edu/Services/Class/PHYS498TBP/spring2002/neuro_0.pdf
- Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv.* Recuperado de https://doi.org/10.48550/arXiv.1708.08296
- Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. (1998). Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Mining and Knowledge Discovery, 2(2), 169–194.* https://doi.org/10.1023/A:1009745219419
- Sharifrazi, D. et al. (2021, febrero 13). Fusion of convolution neural network, support vector machine and Sobel filter for accurate detection of COVID-19 patients using X-ray images. *arXiv*. Recuperado de https://arxiv.org/abs/2102.06883
- Shlens, J. (2014, abril 03). A tutorial on principal component analysis. *arXiv*. Recuperado de https://doi.org/10.48550/arXiv.1404.1100
- SimpliRoute. (2023, enero 10). *Regresión Logística: Qué Es y Cómo Funciona.* Recuperado de https://simpliroute.com/es/blog/regresion-logistica?utm_source=chatgpt.com
- Ultralytics. (s.f.). *Árbol de decisión.* Recuperado el 15 de mayo de 2025, de https://www.ultralytics.com/es/glossary/decision-tree
- van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9(Nov), 2579–2605*. Recuperado de

http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

- Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks, 11(3), 586–600*. Recuperado de https://doi.org/10.1109/72.846731
- Wang, H., et al. (2019, agosto 21). Automatic Machine-Learning-Based Outcome Prediction in Patients With Primary Intracerebral Hemorrhage. *National Library of Medicine*. Recuperado de https://pubmed.ncbi.nlm.nih.gov/31496988/
- Wang, H., Rivenson, Y., Jin, Y., Wei, Z., Gao, R., Günaydın, H., ... & Ozcan, A. (2019). Deep learning enables cross-modality super-resolution in fluorescence microscopy. *Nature Methods, 16(1), 103–110*. Recuperado de https://doi.org/10.1038/s41592-018-0239-0
- Wang, J., Yang, X., Peng, Y., et al. (2023). Diabetic retinopathy detection using LightGBM and transfer learning. *Journal of Healthcare Engineering*, 2021, 1–11. Recuperado de https://www.researchgate.net/publication/374832199_Diabetic_Retinopathy_Detection_Using_Transfer_Learning
- Wattenberg, M., Viégas, F., & Johnson, I. (2016, octubre 13). How to use t-SNE effectively. *Distill.* Recuperado de https://doi.org/10.23915/distill.00002
- Yunus, R. et al. (2014, junio 06). Stroke Prognostication in Patients Treated with Thrombolysis Using Random Forest. *The Open Neuroimaging Journal*. Recuperado de https://openneuroimagingjournal.com/VOLUME/17/ELOCATOR/e18744400298093/FULLTEXT/
- Zhou, Y., Liu, M., Zhang, S., & Tang, X. (2025). Comparison of Machine Learning Models for Predicting Mortality in Acute Ischemic Stroke. *Frontiers in Neurology*, 14, 1167543. Recuperado de https://pmc.ncbi.nlm.nih.gov/articles/PMC11958992/