17 de marzo de 2025

Desarrollo de un modelo predictivo de Machine Learning supervisado para estimar el riesgo de infarto cerebral usando variables clínicas y demográficas.

Daniel Jiménez Valencia
Químico Farmacéutico

Quiero agradecer profundamente a mi buen amigo Fabián Trigo, quien me ayudó a sentar las bases del Machine Learning, dedicándole varias horas de su tiempo a enseñarme con gran calidad pedagógica todo lo necesario antes de comenzar este proyecto.

Resumen

Se desarrolló un modelo predictivo de Machine Learning basado en algoritmos de Regresión Logística y Multilayer Perceptron, ensamblados mediante Voting Classifier. El modelo fue entrenado utilizando datos clínicos y demográficos preprocesados provenientes del conjunto de datos stroke.csv. El preprocesamiento incluyó, entre otras cosas, el balanceo de clases mediante generación de datos sintéticos utilizando CTGAN. El modelo final alcanzó un 93% de sensibilidad, un 92% de F1-score y un valor del área bajo la curva ROC del 97%.

Palabras clave: accidente cerebrovascular, infarto cerebral, inteligencia artificial, machine learning, modelos supervisados.

Índice de contenido

Lista de Abreviaturas

- **ACV:** Accidente cerebrovascular.
- **HTA:** Hipertensión arterial.
- **ML:** Machine learning.
- **IA:** Inteligencia Artificial.
- **RL:** Regresión Logística.
- **KNN:** K-nearest Neighbor.
- **RF:** Random Forest.
- **XGB:** Xtreme Gradient Boost
- **CAT:** CatBoost.
- **RNA:** Redes neuronales artificiales.
- **MLP:** Multilayer Perceptron.
- **ROC-AUC:** Área bajo la curva ROC (Receiver Operating Characteristic).
- **IMC:** Índice de masa corporal.
- **EDA:** Análisis exploratorio de datos.
- **IQR:** Rango Intercuartil.
- **Chi^2:** Chi-Cuadrado
- **MCAR:** Missing completely at random.
- **MAR:** Missing at random.
- **MNAR:** Missing not at random.
- **KDE:** Estimación de densidad de Kernel.
- **CTGAN:** Conditional Tabular GAN
- **VP:** Verdadero positivo.
- **VN**: Verdadero negativo.
- **FP:** Falso positivo.
- **FN:** Falso negativo.

1. Introducción

Un accidente cerebrovascular (ACV) o ictus, es la manifestación de una alteración de un vaso sanguíneo cerebral, generalmente en forma brusca, ya sea por obstrucción o por ruptura. Sus consecuencias en el cerebro suelen ser catastróficas, y los síntomas producidos muy incapacitantes.

Existen dos tipos principales de ictus: el ACV isquémico o infarto cerebral, y el ACV hemorrágico o derrame cerebral.

El infarto cerebral se produce cuando un vaso sanguíneo que irriga sangre al cerebro resulta bloqueado. Esta alteración detiene el suministro normal de sangre, lo que evita que el tejido cerebral reciba oxígeno y nutrientes vitales para su funcionamiento, desencadenando la muerte de las células cerebrales.

La obstrucción de los vasos sanguíneos puede deberse a depósitos de grasa que se acumulan en ellos, o bien, puede deberse a coágulos de sangre. Esto último, puede suceder de dos maneras:

- Se puede formar un coágulo en una arteria que ya está muy estrecha, lo que se denomina accidente cerebrovascular trombótico.
- Un coágulo se puede desprender de otro lugar de los vasos sanguíneos del cerebro, o de alguna parte distinta del cuerpo y trasladarse hasta el cerebro. Esto se denomina embolia cerebral o accidente cerebrovascular embólico.

El derrame cerebral ocurre cuando un vaso sanguíneo de una parte del cerebro se debilita y se rompe, produciéndose una fuga de sangre. Un ACV hemorrágico puede ocurrir a causa de una presión arterial muy alta que haga que los vasos sanguíneos se revienten. Algunas personas tienen defectos en los vasos sanguíneos del cerebro que hacen que esto sea más probable, como por ejemplo un aneurisma. También pueden ocurrir por sobredosis o mal empleo de anticoagulantes, como la warfarina. Un ACV isquémico puede presentar sangrado y convertirse en un ACV hemorrágico.

El principal factor de riesgo para los ACV es la hipertensión arterial (HTA) [1], pero otro factor de riesgo importante es la fibrilación auricular, un tipo de arritmia, ya que los latidos irregulares pueden provocar la formación de coágulos en la aurícula, que luego pueden viajar al cerebro y causar un infarto cerebral. Otras cardiopatías como la enfermedad de las arterias coronarias también aumentan el riesgo de sufrir ACV.

Otros factores de riesgo son la diabetes, antecedentes familiares de la enfermedad, hipercolesterolemia, obesidad y aumento de la edad (especialmente adultos mayores).

Además, existen ciertos hábitos que pueden aumentar el riesgo, tales como el sedentarismo, el alcoholismo, la alimentación rica en grasas saturadas y grasas trans, y el tabaquismo.

El tabaquismo es factor de riesgo por varios motivos. Fumar cigarrillos puede dañar el corazón y los vasos sanguíneos, aumentando el riesgo de sufrir un ACV. La nicotina aumenta la presión arterial. El monóxido de carbono del humo del cigarrillo reduce la cantidad de oxígeno que la sangre puede transportar. La exposición al humo de segunda mano también puede aumentar la probabilidad de sufrir un derrame cerebral.

Anualmente se calcula que aproximadamente 15 millones de personas sufren un ACV, y de estos, unos 5 millones mueren y otros 5 millones quedan con discapacidad severa. La Organización Mundial de la Salud estima que cada 5 segundos ocurre un ACV en la población mundial [2].

"La isquemia puede tardar varias horas en desarrollarse y este tiempo, denominado ventana terapéutica, es un momento clave para evitar o minimizar el daño cerebral" [3]. Esta es una de las cuestiones en donde el Machine Learning (ML) puede ser un gran aporte.

El ML es una subdisciplina de la inteligencia artificial (IA), y surge en la década de los cincuenta como un recurso para emular, computacionalmente, elementos del proceso cognitivo humano a través de reconocimiento de patrones y procesos de toma de decisión. Se basa en un algoritmo que estructura los datos recopilados y, de esta forma, permite al sistema aportar soluciones de manera independiente a los diferentes problemas que se planteen. En el ámbito de la medicina ha sido utilizado para una serie de acciones, como la de aumentar la precisión diagnóstica, predecir la necesidad de ciertas terapias o conocer la efectividad que tendrá un tratamiento en un perfil concreto de un paciente, y, quizás las más importantes en el contexto y motivo de este trabajo, la de emitir diagnósticos que consigan ofrecer datos predictivos sobre ciertas enfermedades.

Los algoritmos de ML se pueden subdividir en dos grandes categorías: aprendizaje supervisado y aprendizaje no supervisado.

En el aprendizaje supervisado, los datos para el entrenamiento incluyen la solución deseada, llamada "etiqueta". Son utilizados para resolver problemas de clasificación, por ejemplo, clasificar imágenes de radiografías en las categorías "sano" o "con neumonía", en el que, para entrenarlo, se

debe incluir si la radiografía presenta neumonía o no con un 1 o un 0. También son usados para resolver problemas de regresión, por ejemplo, predecir el número de bacterias en un cultivo o la concentración plasmática de un fármaco en un tiempo t. En el aprendizaje no supervisado, en cambio, los datos de entrenamiento no incluyen etiquetas, y el algoritmo intentará clasificar o descifrar la información por sí solo. Se utiliza por ejemplo para agrupar información recolectada sobre usuarios de algún servicio y detectar diversas características que tengan en común. A continuación, se resumen algunos algoritmos de aprendizaje supervisado:

Regresión Logística (RL): Es empleado para identificar la relación entre una variable dependiente y una o más variables independientes. La variable dependiente es categórica, lo que significa que hay resultados binarios, como "verdadero" y "falso" o "sí" y "no". Se utiliza, por tanto, principalmente para resolver problemas de clasificación binaria, como puede ser la presencia o no de una enfermedad.

K-nearest Neighbor (KNN): Este algoritmo clasifica los puntos de datos según su proximidad y asociación a otros datos disponibles. Busca calcular la distancia entre puntos de datos, generalmente a través de la distancia euclidiana, y luego asigna una categoría basada en la categoría más frecuente o en el promedio.

Support Vector Machine (SVM): Se aprovecha para problemas de clasificación, construyendo un hiperplano donde la distancia entre dos clases de puntos de datos es máxima. Este hiperplano se conoce como límite de decisión, y separa las clases de puntos de datos (por ejemplo, pápulas frente a pústulas) a ambos lados del plano.

Árbol de decisiones: Se asemeja a un diagrama de flujo. Los datos se dividen en subconjuntos basados en características específicas hasta alcanzar un resultado. El árbol se construye dividiendo iterativamente los datos en función de la característica que mejor separa las clases.

Random Forest (RF): Es una mejora de los árboles de decisión que predice un valor o categoría mediante la combinación de resultados de una serie de árboles en lugar de un solo árbol para mejorar la precisión y reducir la varianza.

Xtreme Gradient Boost (XGB): Construye modelos de árboles de decisión de manera secuencial, corrigiendo los errores del modelo anterior; Se entrena un primer árbol con predicciones iniciales, se calculan los errores entre la predicción y la realidad (residuos), se construye un nuevo árbol que corrige los residuos del anterior, se repite el proceso agregando más árboles, cada uno ajustando

los errores del modelo previo. Finalmente se suman todas las predicciones de los árboles con pesos adecuados para obtener un resultado. Existen más variantes basadas en árboles de decisión, como CatBoost (CAT), especializado para datos tabulares con muchas características categóricas.

Redes Neuronales Artificiales (RNA): Procesan los datos de entrenamiento de entrada al imitar la interconectividad del cerebro humano a través de capas de nodos. Cada nodo se compone de entradas, ponderaciones, un sesgo (umbral) y un resultado. Si ese valor de salida supera un umbral determinado, se activa el nodo, pasando los datos a la siguiente capa de la red. Las redes neuronales aprenden a partir de ajustes basados en la función de pérdida mediante el proceso de descenso gradiente. Multilayer Perceptron (MLP), es un ejemplo de RNA, el cual tiene aplicaciones en clasificación, regresión y procesamiento de datos tabulares.

Las métricas que se usan para medir el poder predictivo de un modelo de ML dependen del tipo de problema que se esté abordando. Las métricas clásicas en problemas de clasificación son la exactitud, la precisión, la sensibilidad, el F1-score y el área bajo la curva ROC (ROC-AUC).

El ROC-AUC es una métrica usada para medir qué tan bien separa el modelo las clases positivas y negativas. Se compone de los conceptos ROC (*Receiver Operating Characteristic*), que es un gráfico que muestra la relación entre la tasa de verdaderos positivos (eje Y) y la tasa de falsos positivos (eje X), y el AUC (*Area Under the Curve*), que entrega un valor entre 0 y 1 que mide el rendimiento global del modelo, siendo un AUC cercano a 1 el indicativo de que el modelo separa bien las clases. AUC entre 0.7-0.8: aceptable; AUC entre 0.8-0.9: bueno; AUC entre 0.9-1.0: excelente.

En el presente trabajo se desarrolla un modelo de ML capaz de predecir ocurrencia de ictus, con la motivación de ayudar en la toma de decisiones en el contexto médico y ayudar a evitar o reducir un posible daño cerebral.

2. Objetivos

Objetivo General:

Desarrollar un modelo de Machine Learning supervisado para predecir el riesgo de ACV en pacientes, integrando datos clínicos y demográficos, con el fin de apoyar la toma de decisiones médicas y la identificación temprana de casos de alto riesgo.

Objetivos Específicos:

- Realizar un análisis exploratorio del conjunto de datos (dataset) disponible, identificando tipos de datos, patrones, distribuciones, correlaciones y posibles sesgos.
- Preprocesar y limpiar los datos mediante técnicas de tratamiento de valores nulos, eliminación de outliers y codificación de variables categóricas.
- Aplicar estrategias de balanceo de clases.
- Seleccionar y evaluar diferentes algoritmos de Machine Learning.
- Validar el modelo con datos reales y evaluar su estabilidad.

3. Metodología

3.1. Conjunto de datos

Para la realización de este proyecto se utilizó un conjunto de datos tabulares en formato CSV llamado stroke.csv, que contiene 5110 muestras correspondientes a pacientes con y sin historial de ACV (stroke) y otras once variables tanto clínicas como demográficas:

- Hipertensión Arterial (con/sin).
- Cardiopatía (con/sin).
- Glicemia promedio (valor numérico en mg/dL).
- Índice de masa corporal (IMC) (valor numérico en kg/m^2).
- Estatus de fumador (fumador, exfumador, no fumador o desconocido).
- Edad (valor numérico en años).
- Estado civil (alguna vez casado/a: sí/no).
- Tipo de residencia (rural o urbana).
- Tipo de trabajo (privado, gubernamental, independiente, niño (sin edad para trabajar) o nunca ha trabajado).
- Género (masculino, femenino, otro).

Además, contiene la variable ID, que corresponde a un número de registro de paciente.

El dataset está disponible en la plataforma Kaggle [4].

3.2. Análisis exploratorio de datos (EDA)

Durante esta etapa:

- Se examinaron los tipos de datos y valores nulos.
- Se analizó la proporción entre pacientes con y sin antecedentes de ACV.
- Se observó la distribución de las variables numéricas y categóricas mediante histogramas, y se evaluó si seguían una distribución normal mediante la prueba de Shapiro-Wilk.
- Se identificaron valores atípicos mediante el método del rango intercuartil (IQR) y diagramas de caja, y se evaluó si estos valores representaban datos válidos o errores de entrada comparándolos con rangos fisiológicamente posibles y valores clínicamente esperados.
- Se visualizó la distribución de las variables numéricas continuas en relación con la variable ACV mediante diagramas de violín.
- Se detectaron qué categorías estaban sobrerrepresentadas.
- Se hicieron análisis de asociación entre las distintas variables y la presencia de ACV.

3.2.1. Pruebas de asociación

Para evaluar la relación entre las variables numéricas edad, IMC y glicemia promedio con la presencia de ACV, se aplicó la prueba U de Mann-Whitney. Además, se utilizaron los coeficientes de correlación de Spearman y Kendall para medir la fuerza y dirección de la asociación entre estas variables y la presencia de ACV.

En el caso de las variables con un fuerte desbalance, como hipertensión y cardiopatía, se usó la prueba exacta de Fisher.

Para analizar la relación entre las variables categóricas y la presencia de ACV, se utilizó la prueba de Chi-Cuadrado (Chi^2).

Para evaluar si la relación entre ACV y las variables que aparentemente estaban relacionadas con ACV eran dependientes de la edad, se ajustó un modelo de regresión logística incluyendo un término de interacción entre la edad y cada variable de interés:

$$logit\big(P(ACV)\big) = \beta 0 + \beta 1 * \text{edad} + \beta 2 * \text{variable} + \beta 3 * (\text{edad} * \text{variable})$$

Donde β3 representa el efecto de moderación de la edad sobre la relación entre la variable y la presencia de ACV.

Se buscaron posibles problemas de multicolinealidad mediante el cálculo del factor de inflación de la varianza (VIF).

3.2.2. Análisis de valores nulos

Se estudió la distribución de los valores faltantes en el dataset para evaluar su impacto en el análisis. Con el objetivo de decidir la estrategia de imputación, se realizaron pruebas estadísticas para determinar el mecanismo subyacente de la ausencia de datos, identificando si estos eran Missing Completely at Random (MCAR), Missing at Random (MAR), o Missing Not at Random (MNAR):

- Prueba MCAR no paramétrica de Jamshidian y Jalal, para evaluar si los valores faltantes seguían un patrón completamente aleatorio.
- Prueba de Kolmogorov-Smirnov, para comparar la distribución de la edad entre registros con y sin valores ausentes.
- Prueba de Mann-Whitney y Chi^2, para analizar la asociación entre los valores ausentes de IMC y otras variables.

3.3. Preprocesamiento de datos

Tanto para el preprocesamiento como para el desarrollo de modelos posterior, se utilizó Python como lenguaje de programación.

Se excluyeron las muestras con edades menores a 5 años.

3.3.1. Tratamiento de valores nulos, desconocidos y atípicos.

Aquellos individuos con valores atípicos que se consideraron errores de entrada fueron eliminados del dataset, en ese sentido se decidió quitar aquellos con IMC mayor a 70 kg/m^2, al tratarse de valores raros desde una perspectiva clínica y estadística.

Con respecto a los individuos categorizados en la variable *estatus de fumador* como "desconocido", se hizo lo siguiente:

- Se asumió que todos los individuos menores de 18 años eran "no fumadores".
- Aquellos con edad igual o mayor a 18 años fueron eliminados del dataset.

Para tratar los valores faltantes de IMC se optó por utilizar una técnica de imputación basada en KNN, el algoritmo KNNImputer [5], con k=5 vecinos. El funcionamiento de KNNImputer, es que, para cada muestra con valores faltantes, se identifican los k vecinos más cercanos en función de las características disponibles y luego imputa los valores nulos utilizando la media o la mediana [6]. En este caso se utilizó la mediana. Las variables utilizadas para predecir el IMC, fueron: *edad, estatus de fumador, hipertensión, cardiopatía, género, tipo de trabajo* y *estado civil.*

Se validó la imputación de valores nulos comparando la distribución de IMC antes y después de la imputación, a través de un gráfico de estimación de densidad de kernel (KDE) y una prueba de Kolmogorov-Smirnov.

3.3.2. Reducción y transformación de variables

Se descartaron aquellas características que, según lo observado en el EDA, no aportaban información predictiva al modelo o que no tenían una relación estadísticamente significativa con la presencia de ACV. Finalmente, solo se consideraron las variables hipertensión, cardiopatía, glicemia promedio, estatus de fumador, IMC y edad.

Se decidió transformar a la variable estatus de fumador en una variable binaria, 1 o 0, dependiendo si el paciente ha fumado o no. Esta nueva variable, entonces, le dio el valor de 1 tanto a los fumadores como a los exfumadores.

Las variables continuas se estratificaron en rangos clínicamente relevantes, tal como se muestran en la **Tabla 1**.

TABLA 1. ESTRATIFICACIÓN DE VARIABLES CONTINUAS.

Rango IMC		Rango Etario		Rango Glicemia Promedio	
Categoría	**Valores (kg/m2)**	**Categoría**	**Valores (años)**	**Categoría**	**Valores (mg/dL)**
Bajo peso	< 18.5	Infante	5 - 12	Baja	< 70
Saludable	18.5 - 24.9	Adolescente	13 - 17	Normal	70 - 100
Sobrepeso	25.0 - 29.9	Adulto Joven	18 - 39	Alta	100 - 125
Obesidad	≥ 30.0	Adulto	40 - 59	Muy Alta	125 - 200
		Anciano Joven	60 - 74	Extremadamente Alta	> 200
		Anciano	≥ 75		

3.3.3. Balanceo de clases

Para abordar la distribución desbalanceada de los participantes entre las clases con y sin ACV, se han generado datos sintéticos de la clase minoritaria utilizando CTGAN (Conditional Tabular GAN), un tipo especializado de red adversaria generativa diseñado específicamente para trabajar con datos tabulares, y que tiene la capacidad de aprender los patrones y relaciones dentro del dataset, para luego generar nuevas muestras sintéticas que preservan las propiedades estadísticas de los datos reales [7].

Antes de utilizar CTGAN para generar datos sintéticos de pacientes con ACV, se redujeron las clases sobrerrepresentadas de cardiopatía, hipertensión y glicemia promedio mediante submuestreo (undersampling).

El procedimiento para el balanceo fue el siguiente:

1) Se creó una copia del dataset preprocesado.
2) Se removieron 2000 muestras que cumplían la doble condición de no tener cardiopatía ni hipertensión.
3) Se removieron 1500 muestras más que cumplían la condición de no tener cardiopatía.
4) Se removieron 200 muestras más que cumplían la condición de no tener hipertensión.
5) Se removieron 50 muestras del rango glicemia promedio normal.
6) Se generaron 3583 muestras sintéticas con ACV utilizando CTGAN (750 épocas).
7) Se ensamblaron las 3583 muestras en el dataset original preprocesado, formando un dataset balanceado.

La selección de las muestras removidas fue realizada de manera aleatoria. El número de muestras a remover fue estimado apoyándose de los histogramas.

3.3.4. Preparación de datos

Antes de entrenar los modelos de ML, se realizaron transformaciones para asegurar que los datos estuvieran en un formato adecuado para el modelado.

Primero, las variables categóricas fueron codificadas mediante *one-hot encoding*, utilizando la función `pandas.get_dummies()` para convertirlas en variables numéricas sin introducir sesgos en la interpretación de los valores.

Luego, se separó la variable objetivo (y), correspondiente a la presencia o ausencia de ACV, del conjunto de características (X).

Finalmente, los datos fueron divididos en conjuntos de entrenamiento y prueba, asignando el 20% de las muestras al conjunto de prueba (`test_size=0.2`). Se aplicó estratificación (`stratify=y`) para garantizar que la proporción de casos con y sin ACV se mantuviera en ambos conjuntos, y se estableció una semilla aleatoria (`random_state=42`) para asegurar la reproducibilidad de los experimentos.

3.4. Construcción del modelo de Machine Learning

3.4.1. Algoritmos experimentados

Se desarrollaron y evaluaron los siguientes modelos de aprendizaje supervisado para predecir la ocurrencia de ACV, y se buscó aquel que presentara el mejor desempeño:

KNN	Árbol de decisiones	RF	XGB	CAT	MLP	Voting Classifier

Voting Classifier [8] es un método de aprendizaje conjunto que ensambla múltiples modelos base para realizar una predicción conjunta. Se exploraron diferentes combinaciones de modelos:

- RF + MLP + CAT + XGB + RL
- XGB + RL
- XGB + MLP
- RL + MLP
- RL + XGB + MLP

Cada uno de estos algoritmos recibió un ajuste óptimo de sus hiperparámetros, los cuales se encontraron utilizando GirdSearchCV [9].

3.4.2. Métricas de evaluación del desempeño

Se midió el desempeño utilizando la precisión, exactitud (accuracy), sensibilidad (recall), F1-score y ROC-AUC.

La exactitud mide la fracción de predicciones correctas sobre el total de predicciones realizadas por el modelo:

$$Exactitud\ (Accuracy) = \frac{VP + VN}{VP + VN + FP + FN}$$

VP: verdadero positivo; VN: verdadero negativo; FP: falso positivo; FN: falso negativo.

La precisión mide cuántos de los predichos como positivos por el modelo son realmente positivos:

$$Precisión\ (Precision) = \frac{VP}{VP + FP}$$

La sensibilidad mide qué tan bien identifica el modelo los casos positivos reales:

$$Sensibilidad\ (Recall) = \frac{VP}{VP + FN}$$

El F1-score combina la precisión y la sensibilidad en una única métrica de rendimiento. Se define como el promedio armónico de precisión y sensibilidad:

$$F1_{score} = 2 * \frac{precisión * sensibilidad}{precisión + sensibilidad}$$

El criterio que se utilizó para escoger al modelo definitivo fue aquel que presentara los mejores resultados de las métricas mencionadas, dándole especial importancia a la sensibilidad, al F1-score y al AUC-ROC. En caso de presentar resultados similares, se prefiere aquel modelo que tardara menos tiempo en entrenar.

3.4.3. Pruebas de sobreentrenamiento y generalización

Con el objetivo de garantizar que el modelo no esté sobreajustado, o, dicho en otras palabras, que no esté simplemente memorizando el conjunto de entrenamiento, sino que tenga una capacidad de generalización adecuada, se llevaron a cabo las siguientes evaluaciones complementarias:

- Comparación de desempeño en conjuntos de entrenamiento y validación, visualizado en curvas de aprendizaje.
- Validación cruzada K-Fold: Se utilizaron diferentes porciones del conjunto de datos para el entrenamiento y la prueba en múltiples iteraciones.
- Evaluación con datos no vistos: Se analizó el comportamiento del modelo ante la entrega de nuevos datos con cambios de distinta magnitud en el valor de las variables.
- Prueba de sensibilidad a errores en variables categóricas mediante permutación aleatoria.

3.5. Uso de IA

La IA ha sido una herramienta fundamental en el desarrollo de este trabajo, especialmente para obtener scripts, ya sea por desconocimiento, o para ahorrar tiempo de programación. Su importancia y rol está ampliamente detallada en el **ANEXO**.

4. Resultados y Discusiones

4.1. EDA

El conjunto de datos cuenta con 5110 filas, correspondiente a los registros de pacientes, y 12 columnas, correspondiente a sus características clínicas y demográficas y a su número de registro ID. Hay un total de 201 valores faltantes, todos ellos dentro de la variable bmi (**Figura 1**).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 5110 non-null   int64
 1   gender             5110 non-null   object
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64
 4   heart_disease      5110 non-null   int64
 5   ever_married       5110 non-null   object
 6   work_type          5110 non-null   object
 7   Residence_type     5110 non-null   object
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object
 11  stroke             5110 non-null   int64
dtypes: float64(3), int64(4), object(5)
```

```
Valores nulos por columna:
id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                201
smoking_status       0
stroke               0
dtype: int64
```

FIGURA 1. TIPOS DE DATOS Y CONTEO DE VALORES NULOS.

El dataset en general está bastante desbalanceado en cuanto a pacientes con o sin hipertensión, cardiopatías y ACV, tal y como se puede observar en las **figuras 2 y 3**. Esto es algo con lo que se tuvo que lidiar antes de ponerse a entrenar los modelos de ML. Además, si no se hace algo al respecto, métricas de evaluación de desempeño como la exactitud, que mide el porcentaje de predicciones correctas sobre el total de predicciones realizadas por el modelo, podrían resultar engañosas, porque el modelo podría simplemente ignorar la clase minoritaria y aun así obtener un gran porcentaje de exactitud.

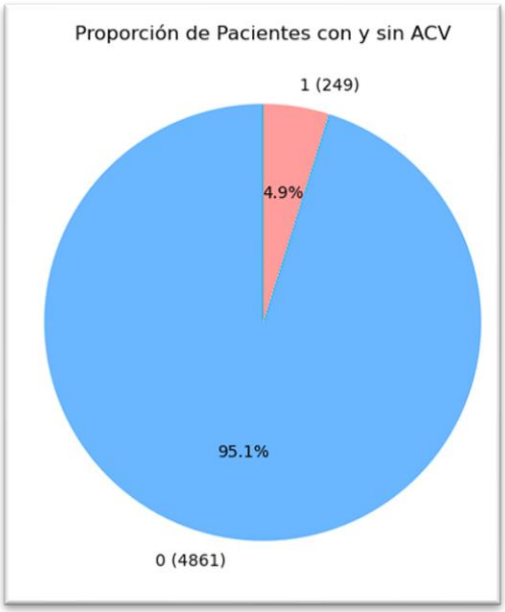


FIGURA 2. PROPORCIÓN DE PACIENTES CON Y SIN ACV.

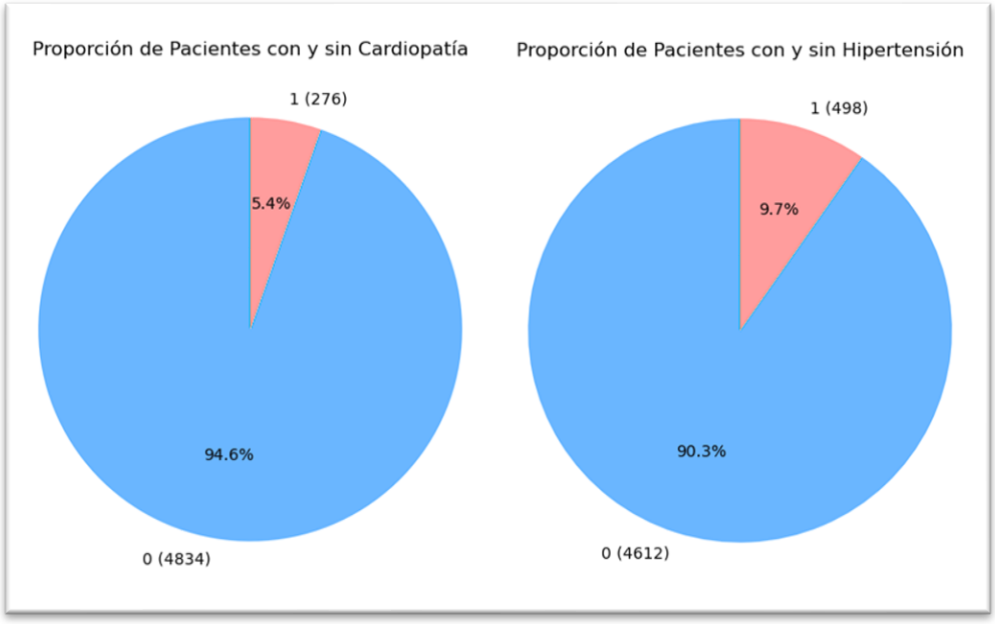


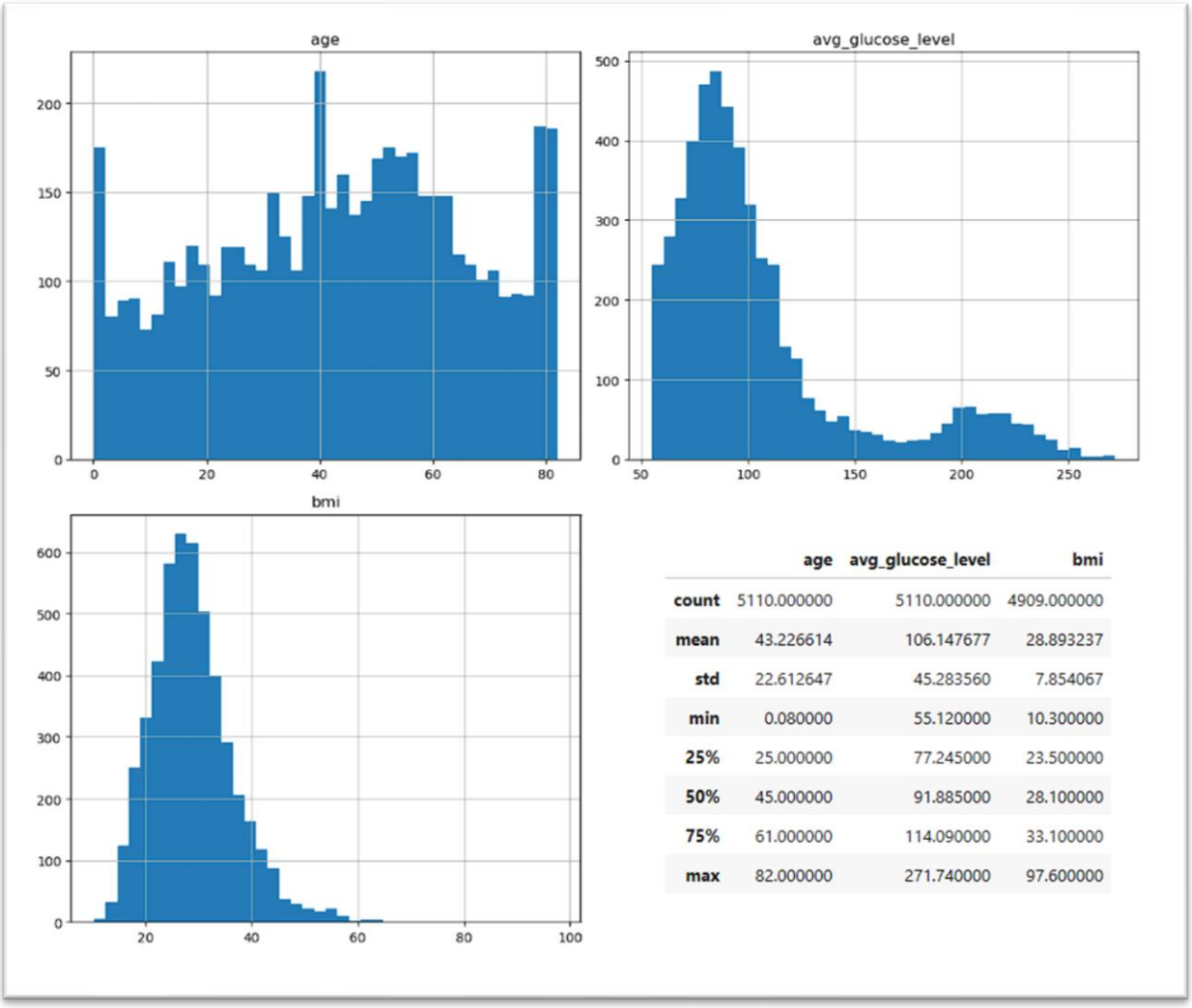
FIGURA 3. PROPORCIÓN DE PACIENTES CON Y SIN CARDIOPATÍAS / HIPERTENSIÓN.

Respecto a las variables numéricas continuas, se puede apreciar gráficamente (**Figura 4**) que no siguen una distribución normal, y así lo demuestra también la prueba de Shapiro-Wilk (**Tabla 2**).

Esto implica que para evaluar la asociación entre estas variables y la presencia o no de ACV, se deben usar pruebas no paramétricas como la prueba U de Mann-Whitney.

TABLA 2. RESULTADOS DE LA PRUEBA DE SHAPIRO-WILK. LAS VARIABLES NO SIGUEN UNA DISTRIBUCIÓN NORMAL.

Variable	Estadístico W	p-value
Edad	0.9672	0.000000
Glicemia promedio	0.9535	0.000000
IMC	0.8059	0.000103



	age	avg_glucose_level	bmi
count	5110.000000	5110.000000	4909.000000
mean	43.226614	106.147677	28.893237
std	22.612647	45.283560	7.854067
min	0.080000	55.120000	10.300000
25%	25.000000	77.245000	23.500000
50%	45.000000	91.885000	28.100000
75%	61.000000	114.090000	33.100000
max	82.000000	271.740000	97.600000

FIGURA 4. HISTOGRAMAS E INFORMACIÓN ESTADÍSTICA DE VARIABLES NUMÉRICAS CONTINUAS. LA MEDIA DE EDAD ES MENOR A LA MEDIANA, LO QUE SUGIERE SESGO HACIA LA IZQUIERDA. POR OTRO LADO, LOS PROMEDIOS DE LOS NIVELES DE GLICEMIA E IMC SON MAYORES QUE SUS MEDIANAS, LO QUE SUGIERE SESGO HACIA LA DERECHA.

En los diagramas de caja (**Figura 5**) se observan un número considerable de outliers en las variables glicemia promedio e IMC. Aunque los que más sorprenden son aquellos con valores de IMC > 70 kg/m^2 (**Figura 6**), al tratarse de valores raros desde el punto de vista clínico, por lo que se decidió tildarlos como valores sospechosos que posiblemente sean errores de entrada.

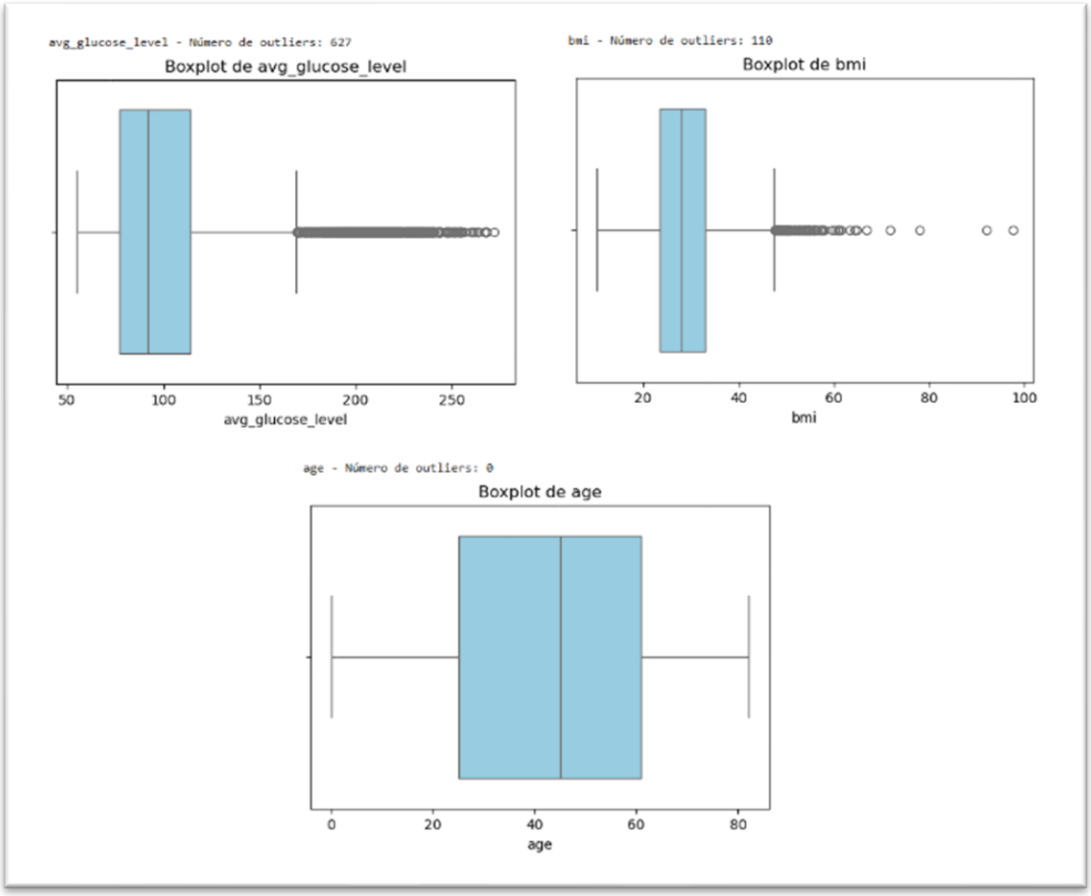


FIGURA 5. DIAGRAMAS DE CAJA Y CONTEO DE OUTLIERS DE VARIABLES CONTINUAS.

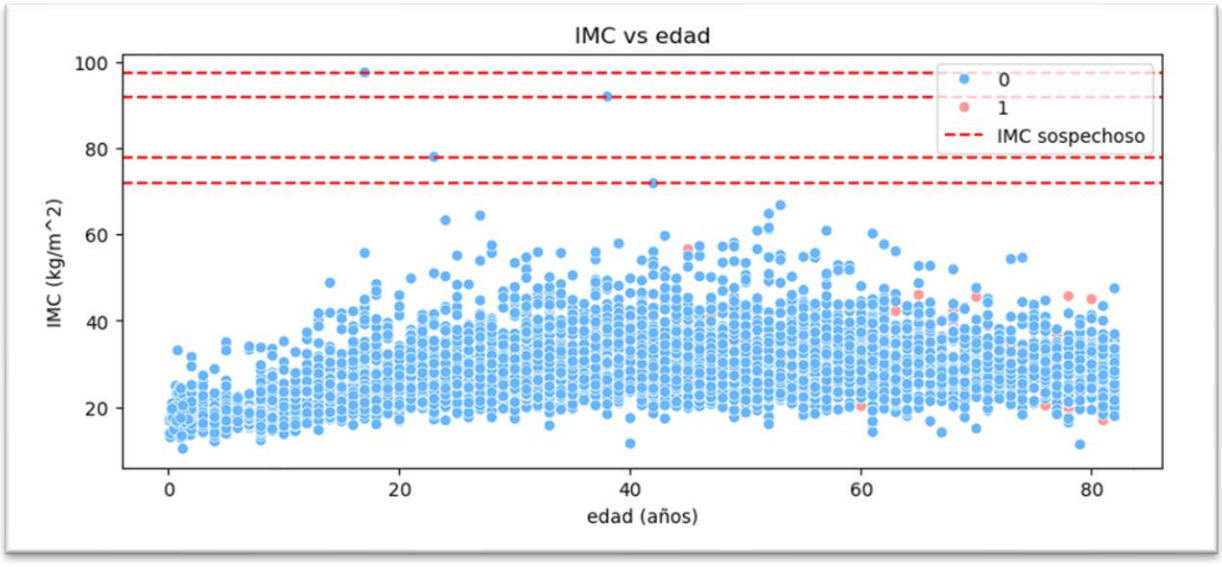


FIGURA 6. DIAGRAMA DE DISPERSIÓN IMC VS EDAD, CON VISUALIZACIÓN DE VALORES RAROS. EL 0 Y 1 SE REFIERE A AUSENCIA O PRESENCIA DE ACV.

En los diagramas de violín (**Figura 7**) se observan diferencias en la distribución de las variables continuas según la presencia de ACV, aunque mucho más notorio en edad y glicemia promedio, cuyas medianas son más altas en el grupo de pacientes con ACV.

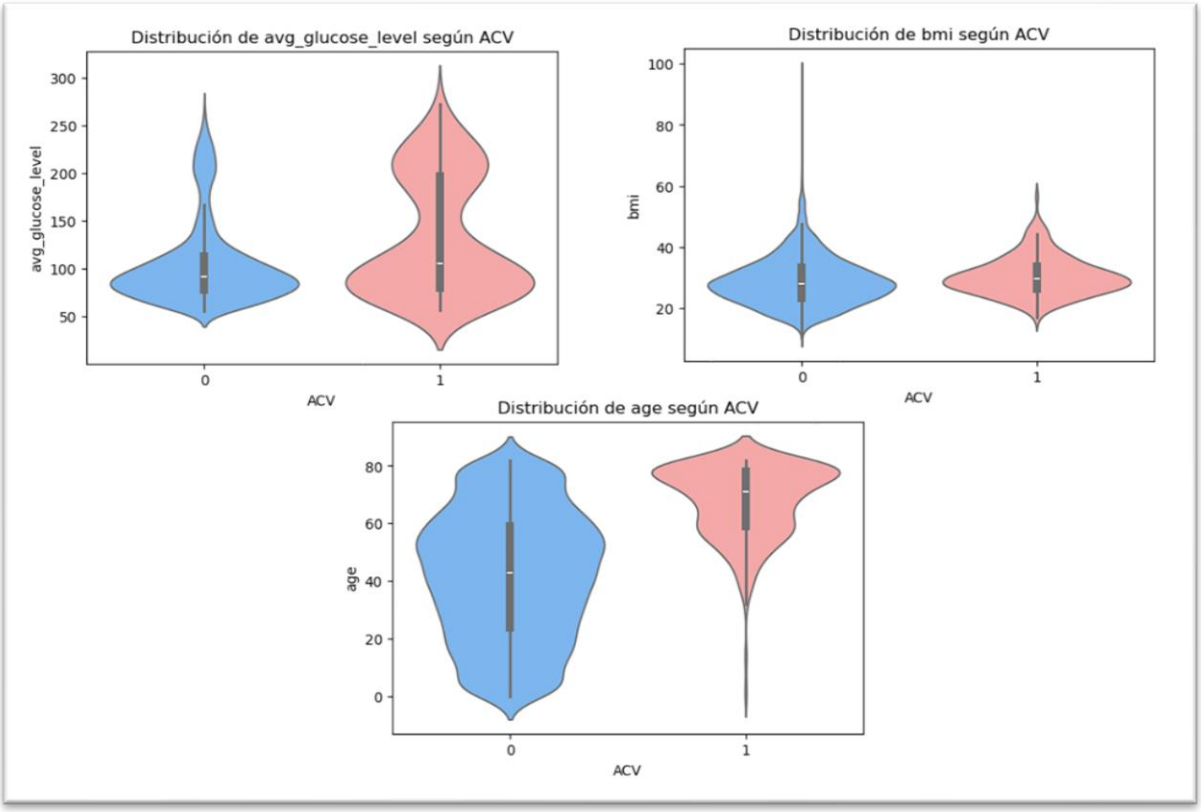


FIGURA 7. VISUALIZACIÓN DE LA DISTRIBUCIÓN DE VARIABLES CONTINUAS MEDIANTE DIAGRAMAS DE VIOLÍN.

4.1.1. Pruebas de Asociación

Los resultados de la prueba U de Mann-Whitney (**Tabla 3**) establecen que existen diferencias significativas en las distribuciones de la edad, IMC y glicemia promedio entre los grupos de pacientes. Esto respalda la hipótesis de que estas variables son factores de riesgo de ACV.

Los coeficientes de correlación de Spearman y Kendall (**Tabla 4**) evidencian que, a pesar de que estas variables muestran correlaciones estadísticamente significativas con ACV, la edad es la variable con mayor correlación, mientras que tanto el IMC como la glicemia promedio tienen correlaciones débiles. Esto sugiere que, si bien el IMC y la glicemia son relevantes, su impacto es menos pronunciado que la edad, la que parece ser un factor crítico en la predicción de ACV.

TABLA 3. RESULTADOS DE LA PRUEBA U DE MANN-WHITNEY.

Variable	Estadístico U	p-value
Edad	161020.5	0.000000
Glicemia promedio	368022.0	0.000000
IMC	413278.5	0.000103

TABLA 4. RESULTADOS DE LAS PRUEBAS DE CORRELACIÓN SPEARMAN Y KENDALL.

Variable	Spearman (r)	p-value	Kendall (τ)	p-value
Edad	0.2351	0.000000	0.1931	0.000000
Glicemia prom.	0.0877	0.000000	0.0716	0.000000
IMC	0.0554	0.000102	0.0454	0.000103

Los resultados de la prueba exacta de Fisher (**Tabla 5**) muestran que tanto la HTA como la cardiopatía están fuertemente asociadas con la presencia de ACV. Los pacientes con HTA tienen aproximadamente 4.44 veces mayor probabilidad de sufrir un ACV en comparación con aquellos sin HTA. Los pacientes con cardiopatía tienen aproximadamente 5.24 veces mayor probabilidad de sufrir un ACV en comparación con aquellos sin cardiopatía.

TABLA 5. RESULTADOS DE LA PRUEBA EXACTA DE FISHER PARA HTA Y CARDIOPATÍA EN RELACIÓN CON ACV.

Variable	Odds Ratio	p-value
Hipertensión	4.4378	0.000000
Cardiopatía	5.2432	0.000000

Las pruebas de Chi-cuadrado (**Tabla 6**) muestran una relación significativa entre la presencia de ACV y el estado de fumador, el estado civil de casado y el tipo de empleo. Sin embargo, las pruebas V de Cramér (**Tabla 7**) sugieren que, aunque existan relaciones significativas entre estas variables y ACV, su relación es débil, por lo que su influencia real sobre el ictus puede ser baja. En el caso del tipo de residencia, prácticamente no tiene relación con la presencia de ACV.

TABLA 6. RESULTADOS DE LA PRUEBA DE CHI² PARA VARIABLES CATEGÓRICAS EN RELACIÓN CON ACV.

Variable	Chi²	p-value	Grados de Libertad
Estado de fumador	34.9435	0.000000	3
Estado civil casado	53.1259	0.000000	1
Tipo de residencia	0.1238	0.724923	1
Tipo de empleo	41.9535	0.000000	4

TABLA 7. RESULTADOS DE LA PRUEBA V DE CRAMÉR.

Variable	Cramér 's V
Estatus de fumador	0.076
Estado civil casado	0.107
Tipo de residencia	0.015
Tipo de empleo	0.098

La aparente relación entre el estado civil y la presencia de ACV está mediada por la edad, ya que, como se observa en el gráfico de la **Figura 8**, hay un cierto valor de edad en donde la proporción de casados se dispara. Y, como ya se observó anteriormente, hay una relación estrecha entre la edad y presencia de ACV.

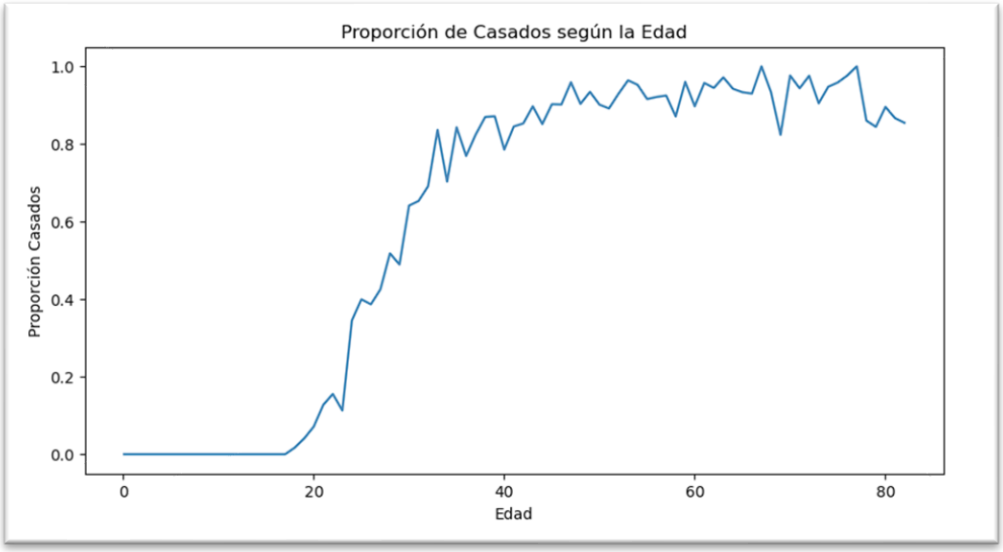


FIGURA 8. GRÁFICO DE PROPORCIÓN DE CASADOS VS EDAD.

Caso similar ocurre con el tipo de empleo, ya que se observa en la **Tabla 8**, que los trabajadores independientes tienen mayor proporción de ACV que el resto de los tipos de empleo. También se observa que aquellos que nunca han trabajado, ninguno ha tenido ACV. Pero, al visualizarlo en el gráfico de proporción de cada tipo de trabajo según la edad (**Figura 9**), se aprecia claramente que la mayor proporción de trabajadores independientes se encuentra en la mayoría de edad, mientras que aquellos que nunca han trabajado se encuentran en el rango etario de adolescente y adulto joven.

TABLA 8. PROPORCIÓN DE ACV SEGÚN TIPO DE EMPLEO.

Tipo de empleo	Sin ACV	Con ACV	Proporción de ACV
Gubernamental	602	28	0.0444
Privado	2684	127	0.0452
Independiente	722	53	0.0684
Nunca ha trabajado	22	0	0.0000
Sin edad para trabajar	670	1	0.0015

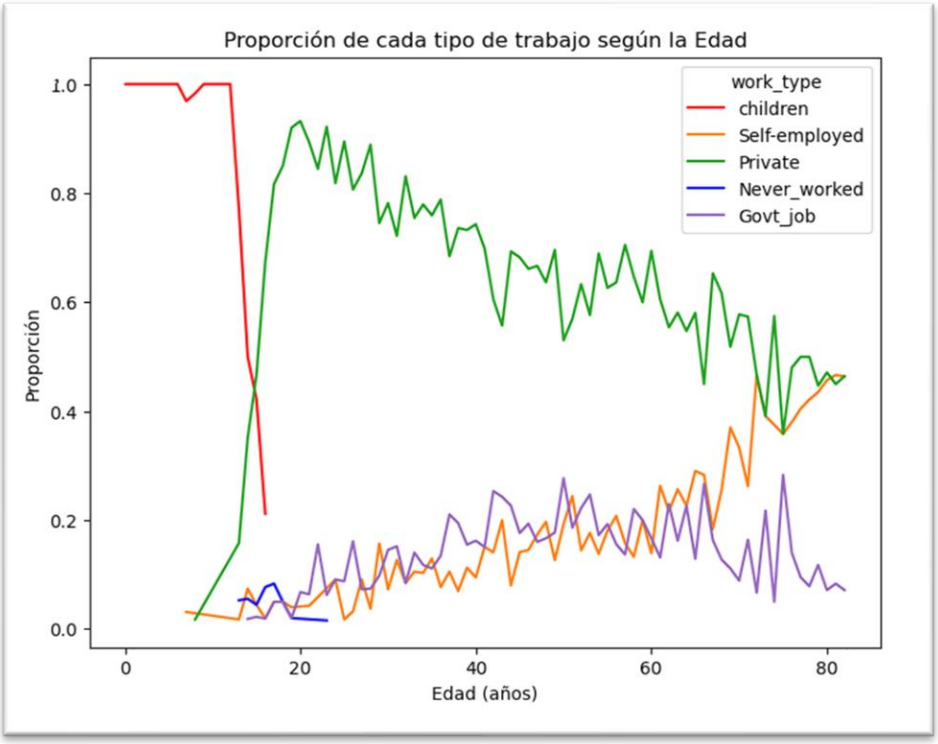


FIGURA 9. PROPORCIÓN DE CADA TIPO DE TRABAJO SEGÚN LA EDAD.

Tras ajustar el modelo de regresión logística incluyendo un término de interacción entre la edad y cada variable de interés (**Tabla 9** y **Tabla 10**), se observa que ni el estado civil, ni el tipo de empleo influyen de manera determinante en el riesgo de ACV, evidenciados por altos p-valores y amplios intervalos de confianza. Por otro lado, se sigue respaldando a la edad como un factor determinante.

TABLA 9. RELACIÓN ENTRE TIPO DE TRABAJO Y ACV LUEGO DE AJUSTAR POR EDAD.

Característica	Odds Ratio	95% IC inferior	95% IC superior	p-value
Trabajador Privado	1.192	0.7976	1.781	0.3915
Trabajador Independiente	0.8066	0.5122	1.270	0.3537
Sin edad para trabajar	3.132	0.6622	14.81	0.1499
Edad	1.083	1.071	1.094	<0.0001

TABLA 10. RELACIÓN ENTRE ESTAR CASADO Y ACV LUEGO DE AJUSTAR POR EDAD.

Característica	Odds Ratio	95% IC inferior	95% IC superior	p-value
Casado	0.8518	0.5574	1.302	0.4584
Edad	1.078	1.068	1.089	<0.0001

En los gráficos de análisis de predicción de probabilidades condicionales mediante regresión logística (**Figura 10**), se observa que la probabilidad de ACV son mayores en los pacientes con hipertensión, cardiopatías y fumadores, independientemente de la edad.

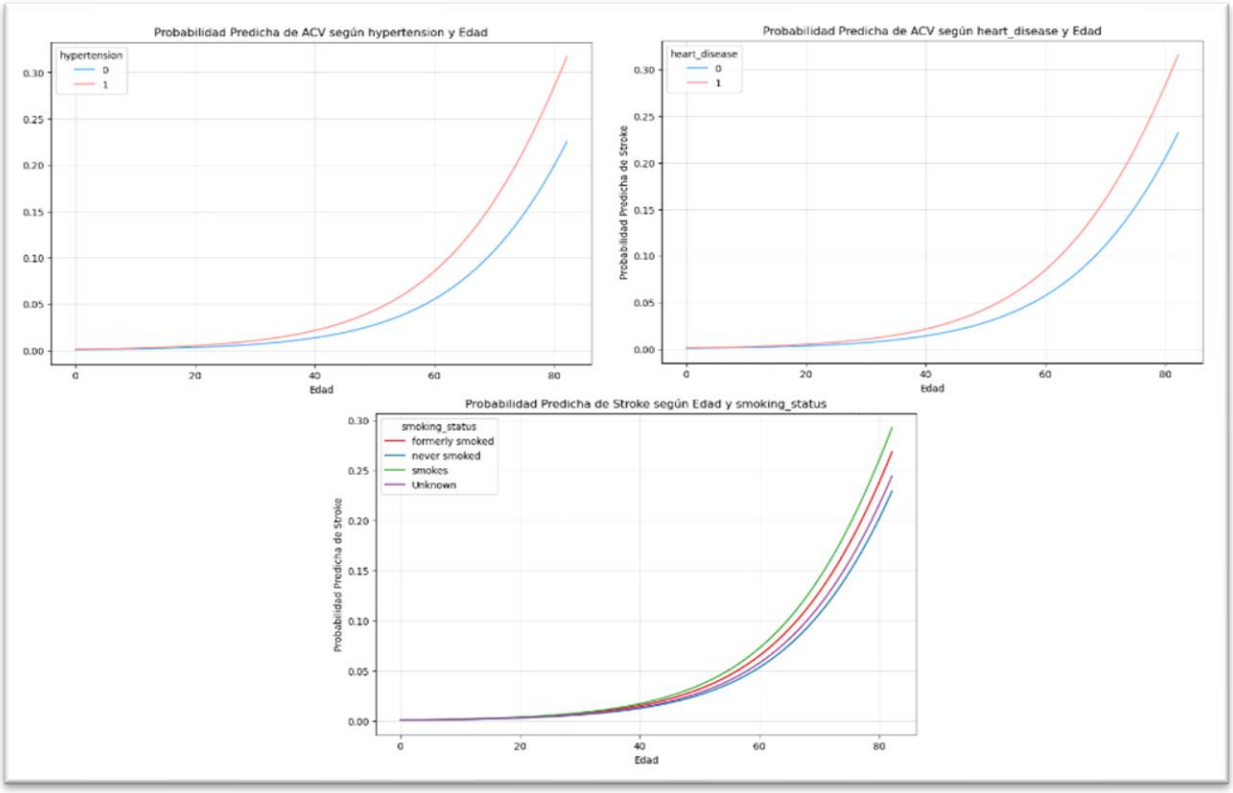


FIGURA 10. GRÁFICOS DE PROBABILIDAD DE ACV SEGÚN VARIABLE VS EDAD.

4.2. Análisis e imputación de valores nulos

Una vez excluidos los pacientes menores a 5 años, existen un total de 187 valores nulos, que corresponden al 4.14% del dataset. 39 de estos valores nulos tienen ACV.

Generalmente, cuando los valores nulos no alcanzan el 5.00% del dataset no se considera una mala decisión eliminarlos. Sin embargo, en este caso, a pesar de que los valores apenas llegan al 4.14%, si se decide eliminarlos, se perderán cerca del 15.7% de los casos con ACV, por lo que se corre riesgo de perder información importante a la hora de querer entrenar los modelos de ML, sobre todo en esta situación en donde se tiene un conjunto de datos sumamente desbalanceado.

Los resultados mostrados en la **Tabla 11**, indican que existen asociaciones significativas entre los valores nulos y variables como la edad y el género. Estos resultados sugieren que los datos faltantes no son completamente aleatorios y probablemente siguen un patrón MAR.

TABLA 11. RESULTADOS DE PRUEBAS ESTADÍSTICAS PARA IDENTIFICAR MCAR, MAR O MNAR.

Prueba	Estadístico/Valor	p-value	Interpretación
Prueba MCAR de Jamshidian y Jalal	0.000000	1.00000	No se rechaza H_0. Los datos faltantes podrían considerarse completamente al azar (MCAR).
Mann-Whitney U (edad entre registros con y sin valores nulos)	610499	0.000000	Existe una diferencia significativa en la distribución de la edad entre registros con y sin valores ausentes.
Kolmogorov-Smirnov (edad entre registros con y sin valores nulos)	0.217791	0.000000	Confirma diferencias significativas en la distribución de edad entre registros con y sin valores nulos, lo que indica que la edad influye en la ausencia de datos.
Chi-Cuadrado (valores nulos en IMC vs género)	9.2704	0.009700	La proporción de valores faltantes en IMC varía significativamente según el género, lo que sugiere que la ausencia de datos está asociada al género.

4.2.1. Validación de la imputación

En la gráfica de KDE, expuesta en la **Figura 11**, se aprecian distribuciones similares de IMC antes y después de la imputación con KNNImputer, lo que indica que no se introdujo sesgo significativo.

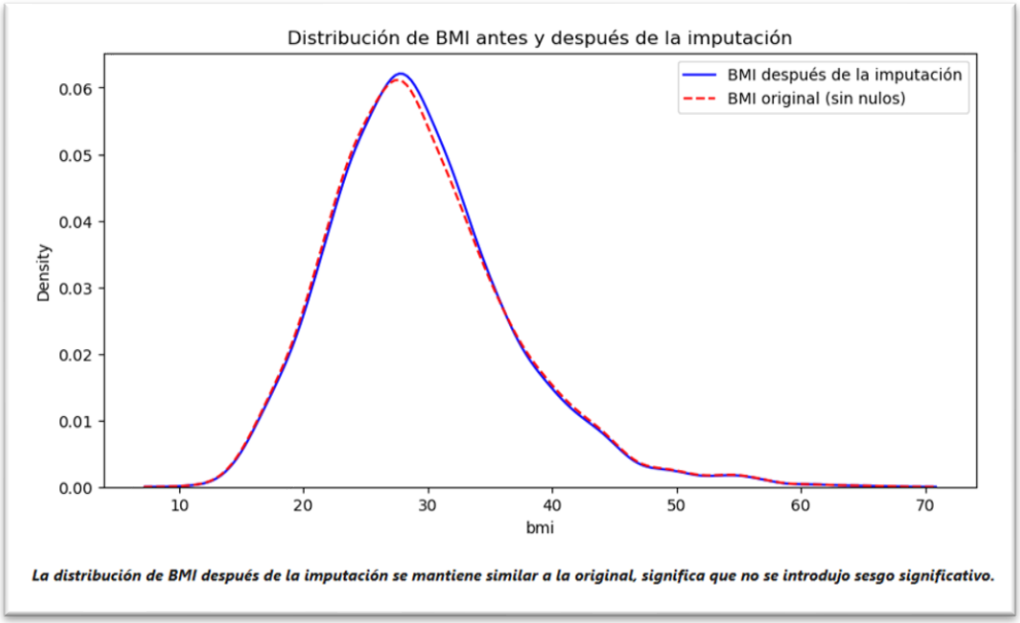


FIGURA 11. GRÁFICA DE ESTIMACIÓN DE DENSIDAD DE KERNEL PARA IMC ANTES Y DESPUÉS DE IMPUTACIÓN.

El resultado de la prueba de Kolmogorov-Smirnov para comparar la distribución de IMC antes y después de la imputación respalda lo observado en la gráfica:

- **Estadístico KS:** 0.0092 y p-valor de 0.9866

4.3. Balanceo de clases

Algunos autores [10, 11], han abordado el problema del desbalanceo de clases del dataset stroke.csv utilizando la técnica SMOTE para generar datos sintéticos de pacientes sin ACV. El problema con esta técnica es que los datos que genera se basan en la interpolación de los datos existentes, y como las variables factores de riesgo (HTA, cardiopatía) también están desbalanceadas, **hay mayor número de casos positivos de ACV dentro de los grupos con ausencia de HTA y cardiopatía**, por lo que SMOTE generará más datos sintéticos con esas características. Ambos trabajos citados muestran excelentes métricas de desempeño, pero ninguno realiza pruebas de sobreentrenamiento o pruebas de robustez, y es probable que sus modelos no generalicen de manera correcta o que incluso estén subestimando el peso de los factores de riesgo mundialmente conocidos. Eso explicaría que en el trabajo de Dritsas E. y Trigka M.[10] se presente una tabla de importancia de características predictivas ("*Table 1. Features importance in the balanced data*") con la variable "ever_married" en segundo lugar de importancia con 9%, y la variable "hypertension" en el último lugar de importancia con solo 0.5%. Además, es probable que haya valores puntuales de edades sobrerrepresentados, y que los modelos predigan excelentemente el riesgo para esas edades puntuales, pero a la hora de probar con variaciones pequeñas de esas edades, sus modelos arrojen porcentajes de riesgo incongruentes. Ese es el motivo por el cual en este trabajo se decidió primero estratificar variables continuas, hacer undersampling en categorías sobrerrepresentadas, y luego generar datos sintéticos, pero con CTGAN.

En las siguientes figuras, se muestran las proporciones de las distintas variables después del preprocesamiento y generación de datos sintéticos con CTGAN. Este conjunto de datos con la variable ACV balanceada fue el utilizado para entrenar los distintos algoritmos de Machine Learning.

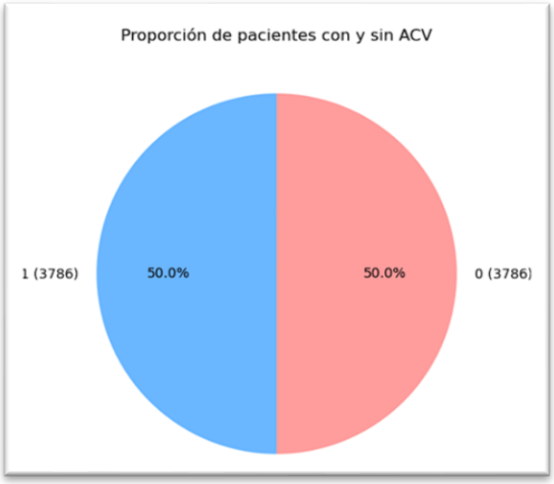


FIGURA 12. PROPORCIÓN DE PACIENTES CON Y SIN ACV DESPUÉS DE BALANCEAR.

FIGURA 13. PROPORCIÓN FINAL DE PACIENTES CON Y SIN HTA.

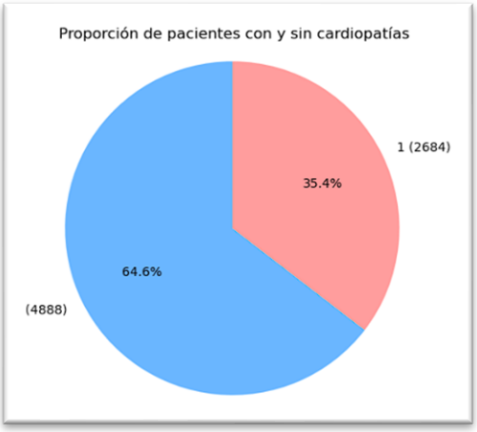


FIGURA 14. PROPORCIÓN FINAL DE PACIENTES CON Y SIN CARDIOPATÍA.

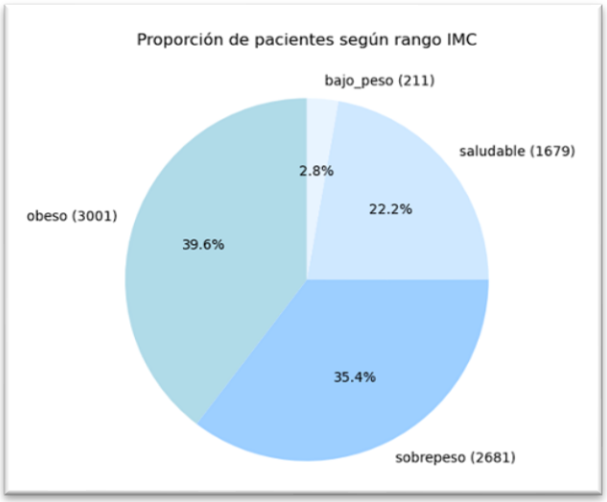


FIGURA 15. PROPORCIÓN FINAL DE PACIENTES SEGÚN RANGO IMC.

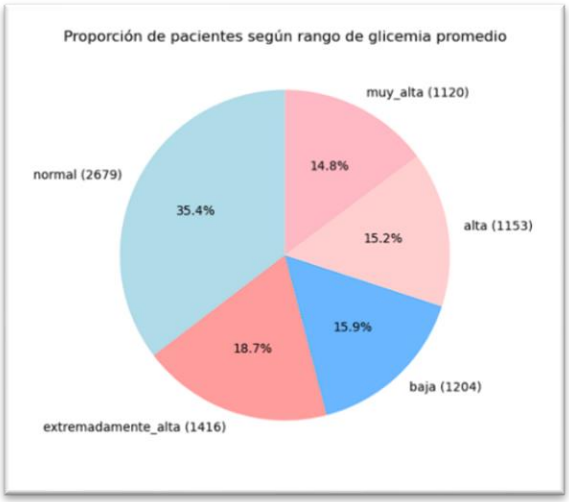


FIGURA 16. PROPORCIÓN FINAL DE PACIENTES SEGÚN RANGO DE GLICEMIA PROMEDIO.

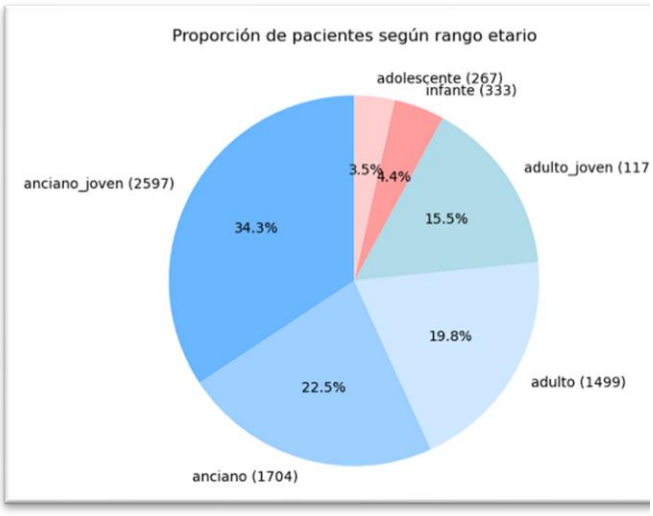


FIGURA 17. PROPORCIÓN FINAL DE PACIENTES SEGÚN RANGO ETARIO.

FIGURA 18. PROPORCIÓN FINAL DE PACIENTES FUMADORES (1) Y NO FUMADORES (0).

4.4. Desempeño de los modelos de Machine Learning

En la **Tabla 12**, se presentan las métricas de desempeño de cada uno de los algoritmos individuales probados. RL destaca por su sensibilidad, XGB por su precisión y ROC-AUC, y MLP por su exactitud y F1-score.

TABLA 12. MÉTRICAS DE DESEMPEÑO DE MODELOS INDIVIDUALES.

Algoritmo	Exactitud	Precisión	Recall	F1-Score	ROC-AUC
Regresión Logística	92.01%	91.84%	**92.21%**	92.02%	96.08%
KNN	90.56%	91.60%	89.30%	90.43%	94.82%
Árbol de Decisiones	91.42%	92.42%	90.22%	91.31%	96.29%
Random Forest	92.08%	92.64%	91.41%	92.02%	96.57%
XGBoost	92.28%	**93.01%**	91.41%	92.21%	**96.89%**
CatBoost	91.88%	92.84%	90.75%	91.78%	96.87%
MLP	**92.34%**	92.79%	91.85%	**92.30%**	96.85%

En la **Tabla 13**, se presentan distintos ensambles mediante Voting Classifier.

TABLA 13. MÉTRICAS DE DESEMPEÑO EN MODELOS ENSAMBLADOS.

Ensambles con Voting Classifier	Exactitud	Precisión	Recall	F1-Score	ROC-AUC
RF + MLP + CAT + XGB + RL	92.28%	92.67%	91.81%	92.24%	**96.71%**
XGB + RL	92.21%	92.32%	92.07%	92.20%	96.58%
XGB + MLP	**92.34%**	**93.14%**	91.41%	92.27%	96.89%
RL + MLP	92.28%	91.99%	**92.60%**	92.30%	96.50%
RL + XGB + MLP	92.34%	92.68%	91.94%	**92.31%**	96.66%

Se escoge el Voting Classifier ensamblando Regresión Logística con MLP como el mejor modelo, debido a que presentó el mejor recall de todos los algoritmos. La sensibilidad es especialmente importante en el contexto de predecir el riesgo de ACV, ya que los falsos negativos serían los errores más costosos. Con un 97% de ROC-AUC, demuestra tener una excelente capacidad para distinguir entre clases positivas y negativas.

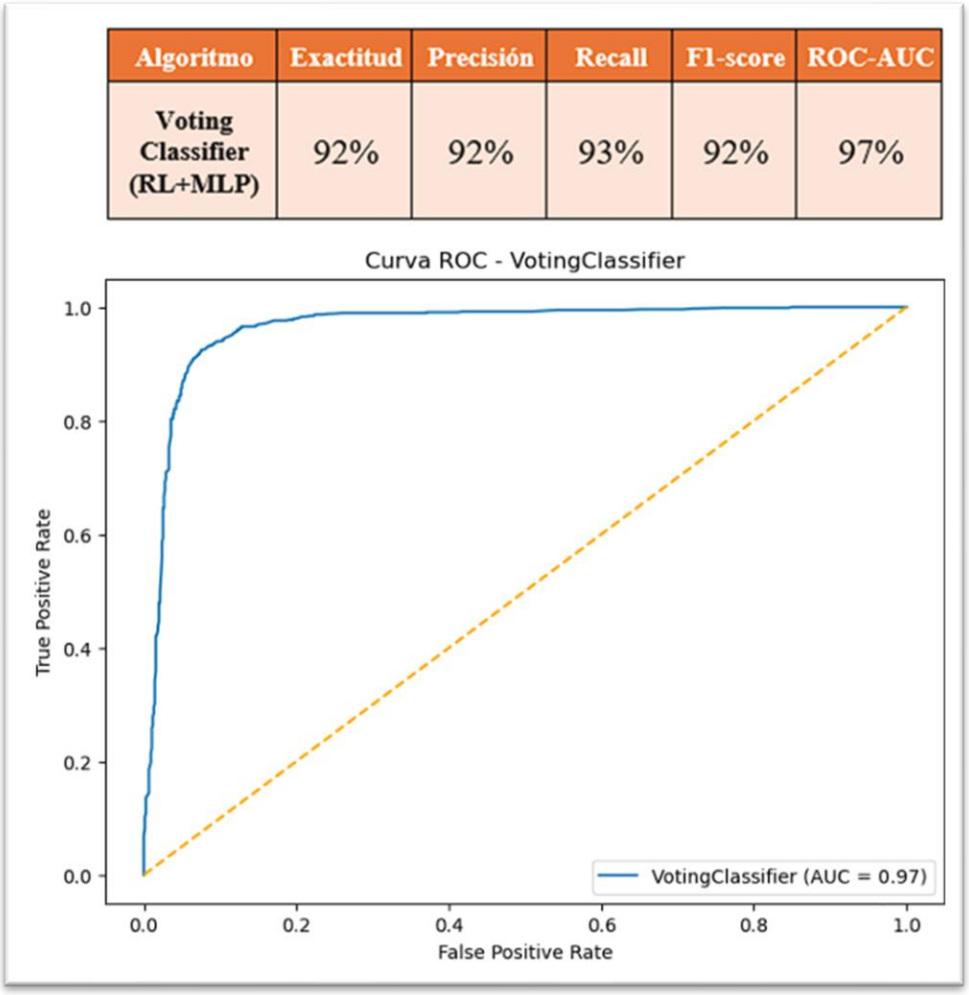


FIGURA 19. MÉTRICAS DE DESEMPEÑO DEL MEJOR MODELO.

- **Hiperparámetros utilizados:**

TABLA 14. HIPERPARÁMETROS UTILIZADOS EN MLP.

Multilayer Perceptron Classifier (MLP)				
hidden_layer_sizes	**activation**	**alpha**	**learning_rate_init**	**random_state**
100	tanh	0.001	0.001	42

TABLA 15. HIPERPARÁMETROS UTILIZADOS EN RL.

Regresión Logística (RL)				
C	**penalty**	**solver**	**max_iter**	**random_state**
10	l2	saga	1000	42

TABLA 16. MODALIDAD DE VOTING CLASSIFIER.

Voting Classifier	
voting	soft

- **Influencia de variables:**

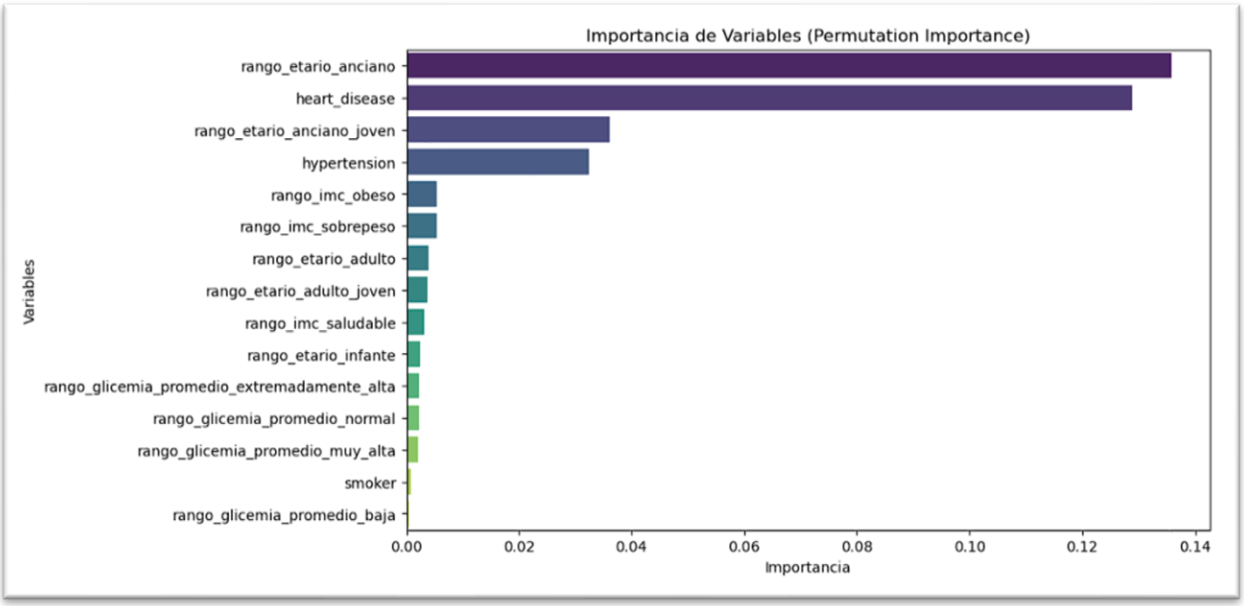


FIGURA 20. PROPORCIÓN DE INFLUENCIA DE LAS VARIABLES. LA EDAD AVANZADA, LAS CARDIOPATÍAS Y LA HIPERTENSIÓN SON LAS VARIABLES QUE APORTAN MÁS INFORMACIÓN PREDICTIVA AL MODELO.

Es probable que el hecho de que cerca del 30% del dataset original fuera del estatus de fumador desconocido, hiciera que no se contara con la cantidad de datos suficientes de fumadores como para que el modelo estimara su importancia como factor de riesgo. Es sugerible obtener una base de datos con mayor muestra de fumadores para futuros modelos. Se planteó la idea de que, en lugar de quitar las muestras con estatus de fumador desconocido con edades igual o mayor a 18 años, se

intentara predecir el estatus fumador verdadero mediante técnicas similares a las utilizadas para imputar valores nulos, sin embargo, se prefirió evitar el riesgo de sesgo.

4.5. Pruebas de generalización, sobreajuste y robustez

- **Comparación de desempeño entre conjunto de entrenamiento y validación:**

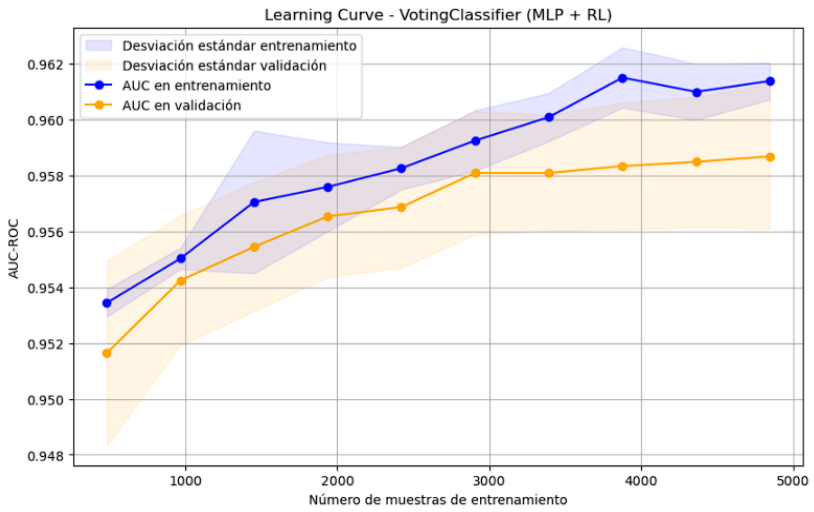


FIGURA 21. CURVA DE APRENDIZAJE PARA ENTRENAMIENTO Y VALIDACIÓN.

La curva de aprendizaje sugiere que el modelo generaliza bien, es estable y no sufre de un sobreajuste significativo, dado que el AUC en entrenamiento y validación convergen y se mantienen cercanos (~0.96).

- **Prueba de Validación cruzada con K-Folds:** 0.9589 ± 0.0062
- **Prueba de robustez ante errores en variables categóricas:**

TABLA 17. VALORES DE AUC DESPUÉS DE APLICAR RUIDO EN VARIABLES CATEGÓRICAS.

Categoría con ruido	Valor de AUC	Diferencia de desempeño
Cardiopatía	0.9597	0.0053
Hipertensión	0.9605	0.0046
Rango etario anciano	0.9574	0.0076
Rango IMC obeso	0.9648	0.0003

Tras aplicar ruido en las variables categóricas más influyentes, el modelo se mantiene con un desempeño similar, lo que demuestra su robustez.

5. Conclusiones

El modelo es eficaz en la detección de ACV. Un 93% de recall asegura que la mayoría de los casos de ACV sean detectados. Su ROC-AUC de 97% muestra que es un clasificador confiable.

La combinación de RL con MLP permite obtener las relaciones lineales entre las variables predictoras y la variable objetivo, al mismo tiempo que aprende patrones no lineales en los datos.

La estrategia de estratificar variables, submuestrear clases sobrerrepresentadas, y luego generar datos sintéticos de la clase minoritaria demuestra ser útil para conseguir un modelo que generalice de buena forma.

Si bien, la variable estatus de fumador no consiguió demostrar la influencia que se esperaba de la literatura científica, se sugiere que puede solucionarse al integrar más muestras de fumadores al conjunto de datos.

Bibliografía

1. Centers for Disease Control and Prevention (CDC). Risk factors for stroke [Internet]. Atlanta: CDC; 2024 [actualizado el 15 de mayo de 2024; consultado el 11 de marzo del 2025]. Disponible en: [CDC].
2. World Health Organization (WHO). Stroke, Cerebrovascular accident [Internet]. Geneva: WHO; [consultado el 11 de marzo del 2025]. Disponible en: [WHO]
3. Clínica Universidad de Navarra. Accidente cerebrovascular o Ictus [Internet]. Pamplona: Clínica Universidad de Navarra; [consultado el 11 de marzo de 2025]. Disponible en: [CUN]
4. Fedesoriano. Stroke Prediction Dataset [Internet]. Kaggle; 2020 [actualizado en 2021; consultado el 11 de marzo de 2025]. Disponible en: [Kaggle]
5. scikit-learn. KNNImputer [Internet]. scikit-learn developers; [consultado el 11 de marzo de 2025]. Disponible en: [scikit-learn]
6. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520-5. https://doi.org/10.1093/bioinformatics/17.6.520
7. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular Data using Conditional GAN. En: Advances in Neural Information Processing Systems (NeurIPS 2019). Vancouver, Canadá; 2019. Disponible en: [NeurIPS]
8. scikit-learn. VotingClassifier [Internet]. scikit-learn developers; [consultado el 11 de marzo del 2025]. Disponible en: [scikit-learn]
9. scikit-learn. GridSearchCV [Internet]. scikit-learn; [consultado el 11 de marzo del 2025]. Disponible en: [scikit-learn]
10. Dritsas, E.; Trigka, M. Stroke Risk Prediction with Machine Learning Techniques. Sensors 2022;22(13):4670. https://doi.org/10.3390/s22134670
11. Wiryaseputra M. Stroke Prediction Using Machine Learning Classification Algorithm. Int J Sci Eng Res. 2022;13(12). Disponible en: [ResearchGate]

APÉNDICE. Notebooks de Preprocesamiento y Modelamiento

Se puede acceder a los notebooks en donde se desarrolla el preprocesamiento y el entrenamiento de modelos en el siguiente repositorio de GitHub: https://github.com/JimenezDValencia/AIctus

ANEXO. Informe de utilización de IA

En este trabajo se han utilizado DeepSeek, versión 3, y ChatGPT-4, versiones o3-mini-high y o1.

DeepSeek-V3, se ha utilizado como un primer acercamiento al desarrollo de modelos de machine learning. De esta IA se obtuvo el workflow y material de estudio recomendado para iniciarse en este campo.

Para tareas de programación, se ha utilizado ChatGPT-4, o3-mini-high. Específicamente, se han obtenido gracias a esta IA, los scripts para generar los gráficos de absolutamente todas las figuras, así como los scripts para realizar el EDA, las pruebas estadísticas, y partes del preprocesamiento claves como la generación de datos con CTGAN, de la cual se detalla en profundidad a continuación.

Cuando se comenzó este proyecto, se utilizaron como referencia los trabajos 10 y 11 citados en la bibliografía. Ambos trabajos utilizaban SMOTE para generar datos sintéticos y, como presentaban buenos resultados, se siguió su metodología. Se obtuvieron métricas de desempeño bastante similares a las suyas. Sin embargo, al probar con pequeñas modificaciones de edad, IMC y niveles de glicemia promedio, el modelo arrojaba resultados incongruentes, tales como mostrar menos riesgo de ACV cuando se pasaba de 80 a 88 mg/dL de glicemia, o una diferencia abismal en porcentaje de riesgo cuando variaba la edad solo un año. Fue en este momento cuando se sospechó que el modelo estaba sobreajustado. Es en este contexto, en donde se consultó a ChatGPT-4, versión o1, quien aportó con la idea de utilizar CTGAN para generar los datos sintéticos, técnica que el autor que redacta este mensaje desconocía. Se utilizó la versión o3-mini-high, para conseguir el "script esqueleto" para generar datos con CTGAN, sin embargo, la intervención humana fue bastante importante, ya que la cantidad de datos generados dependía exclusivamente del procesamiento previo, tales como la imputación de valores nulos, decisiones con respecto al estatus de fumador desconocido, y reducción de clases sobrerrepresentadas según lo observado y concluido en el EDA. La estrategia de balanceo, a gran escala, fue idea del autor.

Cabe destacar que buena parte de los scripts obtenidos con IA, fueron solicitados exclusivamente para ahorrar tiempo, por lo que no fue necesario tener que validar los códigos generados con referencias o literaturas externa, sino que fueron supervisados (y corregidos, si correspondía) por el mismo autor. Para mayor transparencia, se muestra una tabla que clasifica los scripts obtenidos según si fueron solicitados para reducir tiempo, o si fueron solicitados por desconocimiento:

Scripts solicitados para ahorrar tiempo		Scripts solicitados por desconocimiento
EDA	Gráficos de Torta (pie plots)	Prueba de Mann-Whitney
	Histogramas	Correlaciones de Spearman y Kendall
	Gráficos de Caja (box plots)	Prueba Exacta de Fisher
	Gráficos de Violín (violin plots)	Prueba de Chi-Cuadrado
	Line plot de edad vs casados	V de Cramér
	Gráficos de riesgo de ACV según hipertensión, cardiopatía y estado de fumador, considerando edad.	
Preprocesamiento	Tratamiento de la variable estatus de fumador	Prueba de Kolmogorov-Smirnov
	Imputación de valores nulos utilizando KNNImputer	Síntesis de datos con CTGAN, con input de cantidad de muestras según necesidad
	KDE plot	
	Estratificación de variables numéricas	
	Código para remover muestras con ausencia de HTA y/o cardiopatía según necesidad	
Entrenamiento		Búsqueda de hiperparámetros con GridSearchCV
		Configuración y entrenamiento de todos modelos
		Métricas de desempeño
		Pruebas de evaluación de robustez

Los resultados de las pruebas estadísticas cuyos scripts fueron solicitados a la IA por desconocimiento, fueron interpretados 100% por el autor.

La calidad de los scripts desconocidos fue consultada con un experto; Fabián Trigo, licenciado en física con mención en computación científica y doctorando en física de la Universidad de Valparaíso, Chile (https://github.com/fbientrigo - https://www.linkedin.com/in/fabian-trigo/).

No se utilizó IA para la redacción de este trabajo escrito.