UNIVERSIDAD MIGUEL HERNÁNDEZ

Facultad de Ciencias Sociales y Jurídicas de Elche

Grado en Estadística Empresarial

TRABAJO FIN DE GRADO

Ictus bajo la lupa: analizando riesgos con Machine Learning

Curso Académico: 2024 / 2025

Alumno: Carlos Pomares Guirao

Tutora: Maria Asuncion Martinez Mayoral

Elche, junio de 2025

Índice de contenidos

1. Resumen

Este trabajo tiene como objetivo desarrollar y evaluar modelos predictivos capaces de estimar la probabilidad de que un paciente sufra ictus o no, utilizando variables clínicas y demográficas extraídas de una base de datos pública. Se abordan diferentes enfoques de clasificación supervisada, incluyendo regresión logística (con y sin regularización), árbol de decisión, Naïve Bayes y Random Forest. Tras un proceso de preparación de los datos, entrenamiento, optimización de hiperparámetros y validación, se analizan métricas como la exactitud, el AUC y la capacidad para identificar casos positivos, a partir de los cuales se comparan los modelos ajustados. Los resultados muestran que, debido al fuerte desequilibrio de clases, con un grupo minoritario de pacientes que han sufrido ictus, los modelos tienen deficiencias para clasificar correctamente a este grupo. Aunque los modelos desarrollados presentan limitaciones importantes, el modelo **Naïve Bayes** fue el que logró una mejor capacidad para detectar casos de ictus, alcanzando una sensibilidad (recuerdo) del 84% en el conjunto de entrenamiento, aunque con una precisión baja. Su exactitud global fue del 55%, lo que refleja su enfoque centrado en clasificar los casos de ictus, incluso a costa de generar falsos positivos. Esta capacidad lo convierte en el modelo más útil dentro del conjunto evaluado para aplicaciones clínicas donde se prioriza no omitir posibles eventos de ictus.

2.Palabras clave

Las palabras clave son ictus; predicción médica; métricas de evaluación; modelos de clasificación.

3. Contexto

El ictus o accidente cerebrovascular es una de las principales causas de discapacidad y mortalidad en el mundo (*Organización Mundial de la Salud*, 2023), representando una emergencia médica con consecuencias físicas, cognitivas y sociales de gran alcance. Según la Organización Mundial de la Salud, se estima que uno de cada cuatro adultos sufrirá un ictus a lo largo de su vida, lo que pone de relieve la urgencia de desarrollar estrategias efectivas de prevención y detección temprana.

La identificación de los factores de riesgo del ictus ha sido ampliamente estudiada en la literatura médica. Entre los más reconocidos se encuentran la hipertensión arterial, la diabetes, las enfermedades cardiovasculares, el tabaquismo y la edad avanzada (*MedlinePlus*, 2006). Sin embargo, la interacción entre estos factores y su contribución conjunta al riesgo individual continúa siendo un desafío, especialmente en escenarios clínicos donde los recursos diagnósticos son limitados (*Hankey*, 2017).

El presente trabajo se enmarca dentro de esta línea de investigación y utiliza como fuente de datos el conjunto *Stroke Prediction Dataset*, disponible públicamente en la plataforma Kaggle. Esta base de datos recoge información de más de 5000 pacientes,

incluyendo variables demográficas, de salud y estilo de vida, con el objetivo de facilitar el desarrollo de modelos predictivos sobre la ocurrencia del ictus.

Diversos estudios previos en Kaggle han abordado este conjunto de datos (Kse, s.f.; Ferretti, s.f.) desde diferentes perspectivas. Por ejemplo, el análisis exploratorio realizado por Abdullah Kse proporciona una limpieza detallada de los datos y una primera aproximación al modelado mediante algoritmos como Random Forest. Por otro lado, Jacopo Ferretti profundiza en la interpretabilidad de los modelos utilizando técnicas avanzadas como los diagramas SHAP, que cuantifican el impacto de cada variable en la predicción individual del modelo (*Lundberg & Lee*, 2017), y el análisis de dependencia parcial, que representa cómo cambia la predicción del modelo en función de una sola variable (*Molnar*, 2022), aportando una lectura más comprensible para entornos clínicos.

No obstante, la mayoría de estos estudios presentan ciertas carencias: algunos se centran únicamente en la exactitud del modelo sin considerar el impacto del desequilibrio de las clases (los pacientes con ictus son minoritarios frente a los que no lo han padecido); otros se concentran en la aplicación de cierta técnica en particular; pocos incorporan métricas críticas en el ámbito médico, como el recuerdo para ictus.

Este trabajo propone un abordaje amplio, incorporando diversos modelos de clasificación, validación cruzada, evaluación exhaustiva con métricas específicas y un análisis de las posibles variables predictoras. De esta forma, se pretende no solo construir un modelo predictivo eficiente, sino también aportar conocimiento sobre los

factores asociados al ictus, y contribuir al diseño de estrategias preventivas apoyadas en la ciencia de datos.

4. Objetivos

El objetivo de este trabajo es desarrollar y comparar diversos modelos útiles para diagnosticar si un paciente es susceptible de sufrir un ictus en función de diversas variables demográficas, clínicas y de estilo de vida, y que en consecuencia permita diseñar directrices saludables que redunden en la reducción del riesgo de padecerlo. En concreto, se analizarán las variables **hipertensión**, **enfermedad cardíaca**, **tabaquismo**, **índice de masa corporal (BMI)**, **nivel de glucosa en sangre**, **tipo de empleo**, **estado civil**, **género** y **tipo de residencia**.

Los objetivos específicos son:

- Realizar un análisis exploratorio de los datos para describir la distribución de la variable respuesta (ictus) y su relación con las posibles variables predictoras.

- Aplicar técnicas de preprocesamiento de datos: limpieza de valores nulos, transformación de variables, traducción de etiquetas y creación de variables dicotómicas.

- Implementar distintos modelos de clasificación supervisada: regresión logística (con y sin regularización), Naïve Bayes, árbol de decisión y Random Forest.

- Evaluar y comparar los modelos utilizando métricas relevantes como exactitud, precisión, recuerdo, F1-score y AUC.

- Analizar la importancia relativa de las variables predictoras en los modelos generados.

- Aplicar validación cruzada para asegurar la robustez de los resultados.

- Estudiar el comportamiento de los modelos frente al desequilibrio de clases, analizando su impacto en la clasificación de ictus.

- Utilizar técnicas de visualización como curvas ROC, curvas de calibración y gráficos de importancia para interpretar el comportamiento de los modelos.

- Extraer conclusiones clínicas a partir de los resultados y proponer recomendaciones para trabajos futuros.

5. Información disponible

Este trabajo se basa en el análisis del banco de datos *Stroke Prediction Dataset*, disponible en la plataforma pública de Kaggle. Este conjunto de datos fue recopilado y difundido por la empresa Fedesoriano (*Soriano, F. (s.f.), n.d.*), con el objetivo de facilitar investigaciones orientadas a la predicción del riesgo de sufrir un ictus a partir de información demográfica y clínica de pacientes.

El conjunto se presenta en un único archivo en formato CSV, lo cual facilita su descarga y manejo.

La base de datos incluye un total de 5110 observaciones y contempla 12 variables recopiladas.

5.1 Descripción de variables

A continuación, se presenta un listado de las variables utilizadas en el proceso de análisis.

- **ID** → identificador único para cada registro (no se utilizará).
- **GENDER** → variable cualitativa que identifica el género (masculino, femenino u otro)
- **AGE** → variable cuantitativa continua que representa la edad del paciente, y que varía entre 1 y 82 años.
- **HYPERTENSION** → variable binaria relativa a si el paciente tiene (1) o no (0) hipertensión.

- **HEART_DISEASE** → variable binaria relativa a enfermedad coronaria (0= no tiene, 1= sí tiene).
- **EVER_MARRIED** → variable categórica: “Yes” si alguna vez el paciente ha estado casado, “No” si no lo ha estado nunca.
- **WORK_TYPE** → variable cualitativa relativa al tipo de trabajo, con las opciones: “children” (hijos), “govt_job” (trabajo del gobierno), “private” (privado), “self-employed” (autónomo) o “never_worked” (nunca trabajó).
- **RESIDENCE_TYPE** → variable cualitativa sobre el tipo de residencia en la que habita: “rural” (rural) o “urban” (urbano).
- **AVG_GLUCOSA_LEVEL** → variable cuantitativa que mide el nivel promedio de glucosa en sangre, con valores entre 59 y 218.
- **BMI** → variable cuantitativa que mide el índice de masa corporal, con valores entre 12 y 92.
- **SMOKING_STATUS** → variable cualitativa que indica si el sujeto fuma o ha fumado, con las respuestas: “formerly smoked” (antes fumaba), “never smoked” (nunca fumó), “smokes” (fuma) o “unknown” (desconocido).
- **STROKE** → variable binaria con valor 1 si el paciente tuvo un accidente cerebrovascular o valor 0 si el paciente no lo tuvo.

5.2 Procesado de los datos

El procesado de datos es un paso clave al iniciar el tratamiento estadístico. En esta etapa se analiza la existencia de valores atípicos, datos faltantes y posibles inconsistencias, procediendo a eliminar registros atípicos, imputar valores nulos, transformar variables, traducir categorías y generar variables derivadas para facilitar el análisis posterior.

Se observó que en la variable **gender** solo hay un paciente registrado como “other” (las categorías de clasificación son Male/Female), así que se eliminó a este paciente.

Por otro lado, se identificó que la única variable que tiene valores nulos es **BMI**. Debido a que son pocos los valores nulos (201 de 5109), se imputaron los datos faltantes con la mediana de la variable, método reconocido por su robustez frente a la media.

Para llevar a cabo el análisis, se optó por agrupar algunas variables, como el BMI (índice de masa corporal) y el nivel de glucosa en sangre:

- La variable BMI se transforma en una variable dicotómica denominada OBESIDAD.

Creamos una variable dicotómica, a la que denominamos OBESIDAD, para diferenciar pacientes con y sin obesidad, siguiendo el criterio propuesto en MedlinePlus (*Índice De Masa Corporal*, 2025), de donde surgen como posibles valores:

- **0, representando no obesidad** : Si el BMI es menor de 30.
- **1, representando obesidad** : Si el BMI es 30 o más.

- La variable AVG_GLUCOSA_LEVEL se transforma en una variable dicotómica denominada DIABETES.

Creamos una variable dicotómica, DIABETES, para diferenciar pacientes con y sin diabetes, siguiendo el criterio propuesto en MedlinePlus (*Nivel De Glucosa En Sangre*, 2024), de dónde surgen como posibles valores:

- **0, representando no diabetes:** Si el nivel de glucosa está por debajo 126 mg/dL.
- **1, representando diabetes :** Si el nivel de glucosa es igual o superior a 126 **mg/dL.**

Por otro lado, con el objetivo de unificar el manejo de los datos en un solo idioma, hemos procedido a traducir al castellano todos los niveles de respuesta de las variables codificadas en inglés, abordando la consecuente recodificación:

- GENDER
  - **Male** → Hombre
  - **Female** → Mujer
- SMOKING_STATUS
  - **formerly smoked** → Fumó anteriormente
  - **never smoked** → Nunca ha fumado
  - **smokes** → Fumador actual
  - **unknown** → Desconocido

- RESIDENCE_TYPE
  - **urban** → Urbano
  - **rural** → Rural
- WORK_TYPE
  - **Private** → Sector privado
  - **Self_employed** → Autónomo
  - **Govt_job** → Empleado público
  - **Children** → Niño/a
  - **Never_worked** → Nunca trabajó
- EVER_WARRIED
  - **Yes** → Sí
  - **No** → No

6. Metodología

En este proyecto hemos seguido una metodología estructurada en varias fases, abarcando desde el análisis exploratorio inicial hasta la construcción, evaluación y comparación de modelos de aprendizaje automático. A continuación, se detallan las técnicas y herramientas utilizadas en cada etapa.

6.1 Análisis exploratorio

El análisis exploratorio se orienta a describir la variable respuesta (ictus), de tipo categórico, y a analizar su relación con las variables potencialmente predictoras. Cuando las predictoras son categóricas, se han utilizado gráficos de barras apiladas, en los cuales se ha utilizado una codificación por color: **rojo** para representar los casos de pacientes que han sufrido un ictus y **azul** para los pacientes que no han sufrido un ictus.

La relación con variables numéricas se representa con histogramas y gráficos de violín y también con la codificación por color.

6.2 Modelización

Tras el análisis exploratorio y el preprocesamiento de los datos, se procede a la fase de modelización, cuyo objetivo es construir modelos capaces de predecir la clasificación de un paciente como en riesgo o no de padecer ictus, en función de variables demográficas, clínicas y relacionadas con el estilo de vida. Se han explorado diversos enfoques de aprendizaje automático supervisado, con distintos niveles de complejidad e interpretabilidad.

Las variables predictoras consideradas en las modelizaciones incluyen: edad, género, hipertensión, enfermedad cardíaca, estado civil, tipo de empleo, tipo de residencia,

nivel medio de glucosa, índice de masa corporal (BMI), tabaquismo, y las variables derivadas de obesidad y diabetes.

Para poder incluir variables categóricas en los modelos, se procedió a la creación de variables dummy (también conocidas como variables indicadoras), de forma que cada categoría quedó representada como una columna binaria. Además, en aquellos modelos que lo requerían, se realizó la estandarización de las variables numéricas (como edad, glucosa y BMI) para asegurar una escala comparable entre predictores y facilitar la convergencia de algoritmos sensibles a la escala, como la regresión logística regularizada.

Los modelos se entrenaron utilizando un enfoque estándar: división del conjunto de datos en muestras de entrenamiento (80%) y test (20%), utilizando la estratificación por la variable respuesta, dado que la muestra original está muy desequilibrada entre sujetos con ictus y sin ictus; así conseguimos mantener la proporción de casos positivos y negativos en ambas particiones.

Los modelos considerados en este trabajo han sido regresión logística, Naïve Bayes, árbol de decisión y Random Forest, explicados posteriormente.

La evaluación y comparación de los modelos se ha realizado utilizando métricas de clasificación como:

- **Accuracy (Exactitud):** mide la proporción total de predicciones correctas sobre el total de observaciones. Es útil como visión global del rendimiento del modelo, aunque puede resultar engañosa en casos con clases desequilibradas, como el presente.

- **Precisión:** proporción de verdaderos positivos entre todas las predicciones positivas realizadas. Es especialmente relevante cuando se quiere minimizar el número de falsos ictus.

- **Recuerdo (Sensibilidad):** refleja la capacidad del modelo para detectar correctamente los casos de ictus. Es clave en contextos médicos, donde es preferible detectar todos los casos de riesgo aunque haya algunas falsas alarmas.

- **F1-Score:** combina la precisión y el recuerdo en una única métrica armónica, equilibrando ambos aspectos. Resulta útil cuando existe un desequilibrio de clases y se busca un compromiso entre identificar correctamente los ictus y no generar demasiadas falsas alarmas.

- **Matriz de confusión:** representa de forma detallada el número de observaciones correcta e incorrectamente clasificadas como ictus/no ictus. Permite interpretar visualmente los aciertos y errores del modelo, y comprender mejor cómo se distribuyen las predicciones.

- **Curva ROC y AUC (Área Bajo la Curva):** permiten evaluar la capacidad de discriminación del modelo entre clases, analizando su rendimiento a distintos umbrales de decisión. El AUC resume esta información en un valor único entre 0 y 1; cuanto más alto, mayor capacidad predictiva del modelo.

La validación de los modelos se lleva a cabo mediante validación cruzada con **k = 10 particiones**. Este procedimiento consiste en dividir aleatoriamente el conjunto de datos en diez subconjuntos del mismo tamaño. En cada iteración, uno de estos subconjuntos se utiliza como conjunto de validación, mientras que los nueve restantes se emplean para entrenar el modelo. Este proceso se repite diez veces, asegurando que cada subconjunto actúe una vez como conjunto de validación. Finalmente, se calcula la media de las métricas obtenidas en cada iteración, lo que proporciona una estimación robusta del rendimiento del modelo y permite validarlo.

A continuación, se describen los modelos implementados.

6.2.1 Regresión logística

En una primera aproximación se ajustó un modelo de regresión logística para predecir la probabilidad de que una persona sufra un ictus a partir de un conjunto de posibles variables predictoras.

Con el fin de mejorar la generalización del modelo, seleccionar las variables predictoras más relevantes y prevenir el sobreajuste, se evaluaron tres tipos de regularización:

- **Regresión logística con penalización L1 (Lasso)**, que favorece la selección de variables al forzar coeficientes exactamente a cero. Esta técnica modifica la

función de pérdida original de la regresión logística **(log-loss)**, que mide la discrepancia entre las clases reales y las probabilidades predichas. Dicha función se expresa como:

$$Log\ -\ loss\ =-\ \sum[y_i log(p_i)\ +\ (1\ -\ y_i)log(1\ -\ p_i)]$$

donde $\mathbf{y}_i$ representa la clase real del individuo (ictus o no), $\mathbf{p}_i$ es la probabilidad estimada por el modelo de que dicho individuo sufra ictus, y **n** es el número total de observaciones.

Sobre esta función se añade un término de penalización que corresponde a la suma de los valores absolutos de los coeficientes del modelo, dando lugar a la función de pérdida regularizada:

$$Función\ de\ pérdida\ =\ LogLoss\ +\ \lambda\sum|\beta j|$$

Este término penaliza los coeficientes grandes y permite que algunos sean exactamente cero, favoreciendo así la eliminación automática de variables irrelevantes y reduciendo el sobreajuste del modelo.

- **Regresión logística con penalización L2 (Ridge)**, que incorpora una penalización basada en la suma de los cuadrados de los coeficientes:

$$Función\ de\ pérdida = Log - Loss + \lambda\sum\beta j^2$$

A diferencia de Lasso, Ridge no fuerza coeficientes a cero, sino que reduce su magnitud para evitar el sobreajuste, manteniendo todas las variables en el modelo y mejorando su estabilidad.

- **Regresión logística con penalización Elastic Net**, que combina las penalizaciones L1 y L2. La función de pérdida incluye una combinación ponderada de ambas:

$$Función\ de\ pérdida = Log - Loss + \lambda(\sum|\beta j| + (1 - \alpha)\sum\beta j^2)$$

Este enfoque permite ajustar el modelo con la flexibilidad de Lasso (selección de variables) y la robustez de Ridge (coeficientes más estables), siendo especialmente útil en situaciones con alta correlación entre predictores o gran número de variables.

Para obtener los valores óptimos de los términos de penalización se realizó una **optimización de hiperparámetros** (los relativos a la función de pérdidas en la regularización) mediante la técnica de búsqueda en grid con validación cruzada.

Una vez obtenidos los modelos óptimos en cada tipo de regularización, se comparan los tres ajustes regularizados, junto con el ajuste sin regularización y se opta por la mejor de ellas en términos del menor AIC, que equilibra ajuste y complejidad y que se calcula a partir de la verosimilitud del modelo, según la fórmula:

$$AIC = 2k - 2ln(L)$$

Siendo **k** el número de parámetros estimados en el modelo y **L** la verosimilitud del modelo.

Se grafican sus coeficientes para identificar qué variables resultan más influyentes.

Finalmente, se evaluó el rendimiento sobre el conjunto de test con las métricas de clasificación habituales y se representó gráficamente la curva ROC del modelo óptimo. Dado que este tipo de curva ya ha sido introducido previamente, se utiliza aquí como apoyo visual para analizar la capacidad del modelo en distintos umbrales. La implementación y el cálculo del AUC se realizaron utilizando las funciones disponibles en la biblioteca *scikit-learn* (Pedregosa et al., 2011).

6.2.2 Naïve Bayes

Este método se ha seleccionado por su **simplicidad, rapidez y buena capacidad predictiva** en contextos donde las variables predictoras continuas pueden aproximarse razonablemente bien a una distribución normal.

A diferencia de otras variantes como el Naïve Bayes Bernoulli o Multinomial, el **Naïve Bayes Gaussiano** resulta especialmente adecuado en este caso, ya que las variables predictoras utilizadas en el modelo (como la glucosa, el índice de masa corporal o la edad) son de naturaleza continua.

El algoritmo de **Naïve Bayes** se basa en el **teorema de Bayes**, una fórmula probabilística que permite calcular la probabilidad de que ocurra un evento, dado que ya ha ocurrido otro. Este enfoque parte de la siguiente expresión:

$$P(C|X) = P(X|C) \cdot P(C) / P(X)$$

donde:

- P(C|X) es la **probabilidad posterior** de que un paciente haya sufrido ictus dado un conjunto de características X.
- P(X|C) es la **verosimilitud**: la probabilidad de observar las características X dado que se pertenece a la clase C (haber sufrido ictus).
- P(C) es la **probabilidad previa** de la clase C.
- P(X) es la probabilidad de observar el conjunto de características X (independiente de la clase).

Durante la predicción, el modelo evalúa la probabilidad de que un nuevo paciente pertenezca a cada clase y asigna aquella con mayor probabilidad posterior. El resultado es un clasificador rápido, eficiente y sorprendentemente competitivo en muchos

contextos, especialmente cuando se prioriza el **recuerdo** (detección de positivos) por encima de la precisión.

Tras el entrenamiento del modelo sobre el conjunto de entrenamiento, se evaluó su rendimiento sobre el conjunto de prueba. Se calculó la matriz de confusión y las métricas de clasificación (precisión, *recuerdo*, F1-score) por clase. Además, se generaron la curva ROC y la curva de aprendizaje.

La curva de aprendizaje permite visualizar cómo varía el rendimiento del modelo al aumentar el tamaño de la muestra de entrenamiento, ayudando a detectar posibles problemas de sobreajuste o infraajuste (*Raschka*, 2020).

6.2.3 Árbol de decisión

En este apartado se emplea un **modelo de árbol de decisión** para la clasificación. Este tipo de modelo tiene la ventaja de ser altamente interpretable, porque proporciona un árbol en el que sólo aparecen las variables que ayudan a la predicción de la respuesta. Por contra, no proporciona información sobre el grado en el que afectan a la predicción, como sí hace el modelo de regresión a través de los coeficientes estimados.

El algoritmo de **árbol de decisión** para la clasificación divide el espacio de datos en regiones homogéneas a través de una estructura jerárquica de decisiones. El objetivo es crear un árbol en el que cada **nodo interno** representa una condición sobre una variable predictora, cada **rama** representa el resultado de esa condición, y cada **hoja** representa una predicción final (en este caso, ictus o no ictus).

El proceso de construcción del árbol se realiza de forma **recursiva**, mediante el algoritmo conocido como **ID3** (*Quinlan*, 1986) o una de sus variantes como **CART** (*Breiman et al.*, 1984), siguiendo estos pasos:

1. **Selección de la mejor variable para dividir**: en cada nodo se selecciona la variable que mejor separa las clases y los valores/clases frontera para realizar la división. Para ello, se evalúa la "impureza" del nodo, es decir, el grado de mezcla entre clases. Las métricas más utilizadas para medir la impureza son:

    - **Índice Gini**:

$$Gini = 1 - \sum Pi^2$$

    - **Entropía**:

$$Entropía = - \sum Pi * log^2(Pi)$$

    donde Pi es la proporción de elementos de la clase i en el nodo.

2. **División del nodo**: se selecciona la división que maximiza la reducción de impureza, creando dos nuevos nodos hijos. Este proceso continúa de forma recursiva.

3. **Criterio de parada**: la división se puede detener cuando se alcanza cierta profundidad máxima, el número mínimo de muestras por nodo, o cuando los nodos son puros (solo contienen observaciones de una clase). Estas restricciones puede ajustarlas el usuario.

El árbol resultante es fácil de interpretar: cada camino desde la raíz hasta una hoja constituye una **regla de decisión** que combina condiciones lógicas sobre las variables predictoras.

Inicialmente, se entrenó un Árbol de Decisión utilizando los parámetros por defecto del algoritmo, lo cual dio lugar a un modelo sobreajustado, con rendimiento perfecto sobre el conjunto de entrenamiento, pero pobre sobre el conjunto de prueba.

La técnica de poda por complejidad (Cost-Complexity Pruning) se basa en introducir un criterio de penalización que favorece árboles más simples, con el objetivo de mejorar su capacidad de generalización y evitar el sobreajuste. Para ello, se define una función de pérdida regularizada de la siguiente forma:

$$R\alpha(T) = R(T) + \alpha \cdot |T|$$

donde:

- R(T) es el error de clasificación del árbol T,
- |T| es el número de nodos terminales (hojas),

- α es el parámetro de complejidad que penaliza la complejidad estructural del árbol.

El parámetro α actúa como un regulador del equilibrio entre ajuste y simplicidad: a mayor valor de α, más penalizados estarán los árboles complejos, favoreciendo estructuras más pequeñas que conserven un rendimiento adecuado.

Con el objetivo de corregir el sobreajuste observado en el modelo inicial, se aplicó esta técnica explorando distintos valores de α, y seleccionando aquel que minimizó el error de validación, es decir, el que ofreció el mejor compromiso entre capacidad predictiva y complejidad estructural del árbol.

A continuación, se reportaron las métricas de clasificación en entrenamiento y test, así como el número de nodos terminales del árbol final y su profundidad.

Por último, se ha generado una **matriz de importancia de variables** a partir del modelo podado. Esta matriz permite identificar qué predictores han tenido mayor relevancia en la construcción del árbol. La importancia de una variable en un árbol de decisión se define como la cantidad total de reducción de impureza (como el índice Gini o la entropía) que esa variable aporta al dividir los nodos del árbol. Cuanto más utilice el modelo una variable para realizar divisiones que separen las clases, mayor será su importancia.

6.2.4 Random Forest

En esta sección se desarrolló un modelo basado en el algoritmo de **Random Forest**, una técnica de ensamblado que construye múltiples árboles de decisión y combina sus predicciones para mejorar la capacidad predictiva y la estabilidad del modelo. A diferencia de un único árbol, Random Forest reduce el sobreajuste mediante el entrenamiento de varios árboles sobre subconjuntos aleatorios del conjunto de datos, añadiendo también en el algoritmo la aleatorización de las variables que intervienen en la partición de cada nodo.

Con el fin de optimizar el modelo, en términos de la profundidad del árbol, el número de predictores por división, el número de árboles en el bosque, e incluso el criterio de impureza a utilizar (Gini o entropía), se llevó a cabo una **búsqueda exhaustiva de hiperparámetros** utilizando la técnica de búsqueda en grid.

El rendimiento de cada combinación se evaluó mediante el **Out-of-Bag Score (OOB)**, y se seleccionó la que mejor rendimiento ofrecía. El OOB es una métrica de validación interna propia de los modelos de Random Forest, que permite estimar el rendimiento del modelo sin necesidad de recurrir a un conjunto de validación externa (*Breiman*, 2001).

Durante el entrenamiento, cada árbol del bosque se construye a partir de una muestra aleatoria con reemplazo del conjunto de entrenamiento (técnica conocida como **bootstrap**). El modelo utiliza estas muestras OOB para **evaluar el rendimiento predictivo**: cada observación se predice con aquellos árboles en los que no fue

utilizada para el entrenamiento. Posteriormente, se calcula la precisión promedio de todas estas predicciones, proporcionando así una estimación del rendimiento general del modelo.

Por último, se calculan las métricas de clasificación y la matriz de confusión del modelo final.

6.3 Software y hardware

El análisis y modelización han sido realizados utilizando el lenguaje de programación **Python**, en su versión **3.11**.

Hardware utilizado

- **Procesador**: Intel Core i7-10750H CPU @ 2.60GHz
- **Sistema operativo**: Windows 11 Pro

No se han presentado necesidades específicas de procesamiento, dado que el volumen de datos y la complejidad de los modelos permiten su ejecución en un ordenador de gama media-alta sin problemas de rendimiento.

Librerías utilizadas

Se han empleado las siguientes librerías y módulos de Python:

- **pandas**
  Utilizada para la carga, manipulación y gestión de los datos tabulares.

- **numpy**
  Aplicada para la realización de operaciones numéricas y manejo de arrays.

- **matplotlib**
  Utilizada para la generación de gráficos básicos, como los gráficos de barras y el gráfico de violín.

- **seaborn**
  Empleada para crear gráficos estadísticos más elaborados y estéticos, complementando a matplotlib.

- **scikit-learn**
  Principalmente utilizada para modelos de aprendizaje automático, proporciona funcionalidades específicas para:
  - División de los datos en conjuntos de entrenamiento y test.
  - Construcción de modelos de regresión logística con diferentes técnicas de regularización (L1, Elastic Net).
  - Búsqueda de hiperparámetros mediante validación cruzada (GridSearchCV).
  - Cálculo y visualización de la matriz de confusión.

7. Resultados

En esta sección se presentan los resultados obtenidos tras aplicar las diferentes metodologías descritas previamente. Se analizan, en primer lugar, las relaciones observadas entre las variables explicativas y la variable respuesta mediante técnicas de análisis exploratorio. A continuación, se detallan los resultados de los modelos de clasificación implementados, evaluando su rendimiento a través de métricas relevantes como la exactitud, la sensibilidad, la precisión, el F1-score y el área bajo la curva ROC (AUC).

7.1 Análisis exploratorio

Para comprender la distribución de la variable objetivo del estudio,en la Figura 1 se ha representado gráficamente el número de observaciones con y sin ictus.

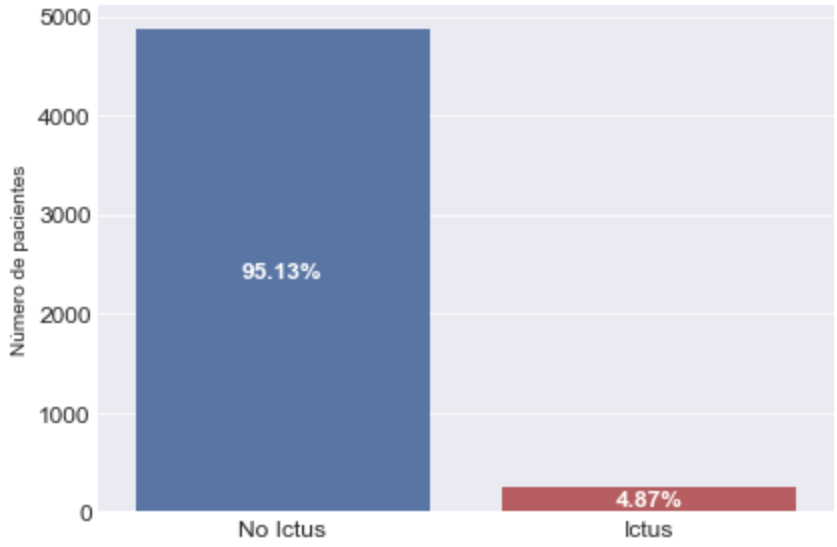

Figura 1. *Número de pacientes con ictus y sin ictus.*

Como se observa en la Figura 1, existe un fuerte desequilibrio en la muestra, siendo la mayoría de los casos personas que no han sufrido un ictus (95.13%).

Procedemos a mostrar gráficamente las relaciones que hay entre cada variable y el hecho de que se sufra un ictus:

En la Figura 2 se muestra la distribución de edades para los pacientes que han sufrido ictus (a la derecha) y los que no (a la izquierda).

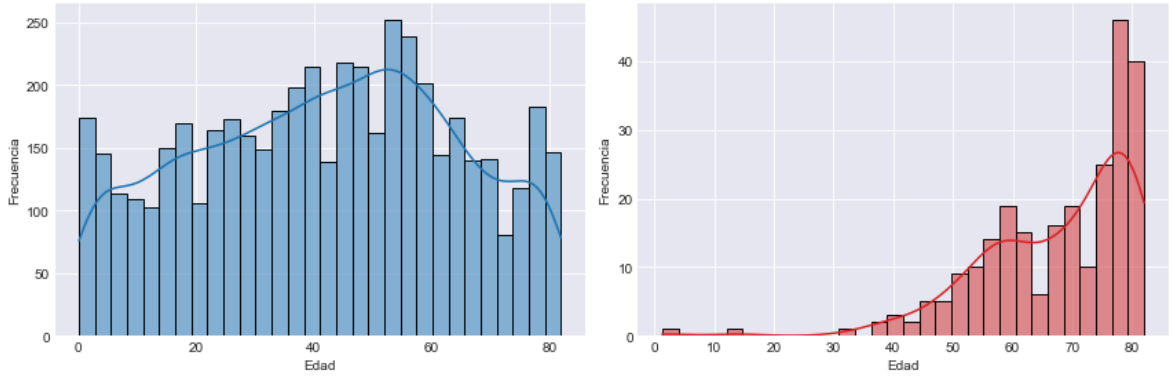

Figura 2. *Distribución de la edad de los pacientes que han sufrido ictus (a la derecha) y los que no (a la izquierda).*

Gracias a estos gráficos (Figura 2) podemos apreciar que hay una cierta relación entre edad y padecer ictus: si bien los pacientes que no padecen ictus están distribuidos prácticamente sobre todo el rango de edad entre 0 y 80 años, el 75% de los pacientes que han sufrido ictus tienen una edad superior a 30 años"

En la Figura 3 se muestra la proporción de pacientes que han sufrido un ictus y los que no, diferenciados por género.

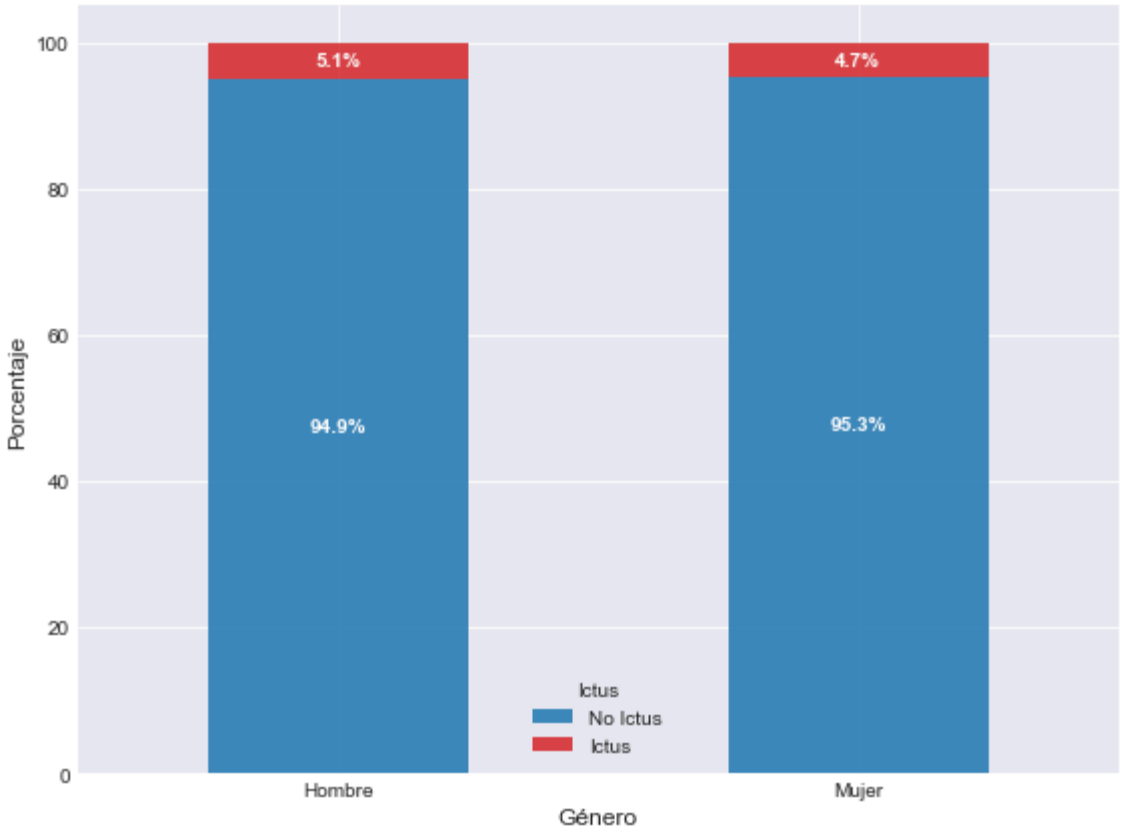

Figura 3. *Distribución porcentual del ictus según el género del paciente.*

Podemos observar en la Figura 3 que el porcentaje de personas que han sufrido un ictus es muy similar entre hombres (5.1%) y mujeres (4.7%). Esto sugiere que el género no parece tener mucha influencia en la probabilidad de padecer un ictus, ya que las diferencias entre ambos grupos son mínimas.

En la Figura 4 se muestra la proporción de pacientes que han sufrido un ictus y los que no, teniendo en cuenta si presentan antecedentes de hipertensión o no.

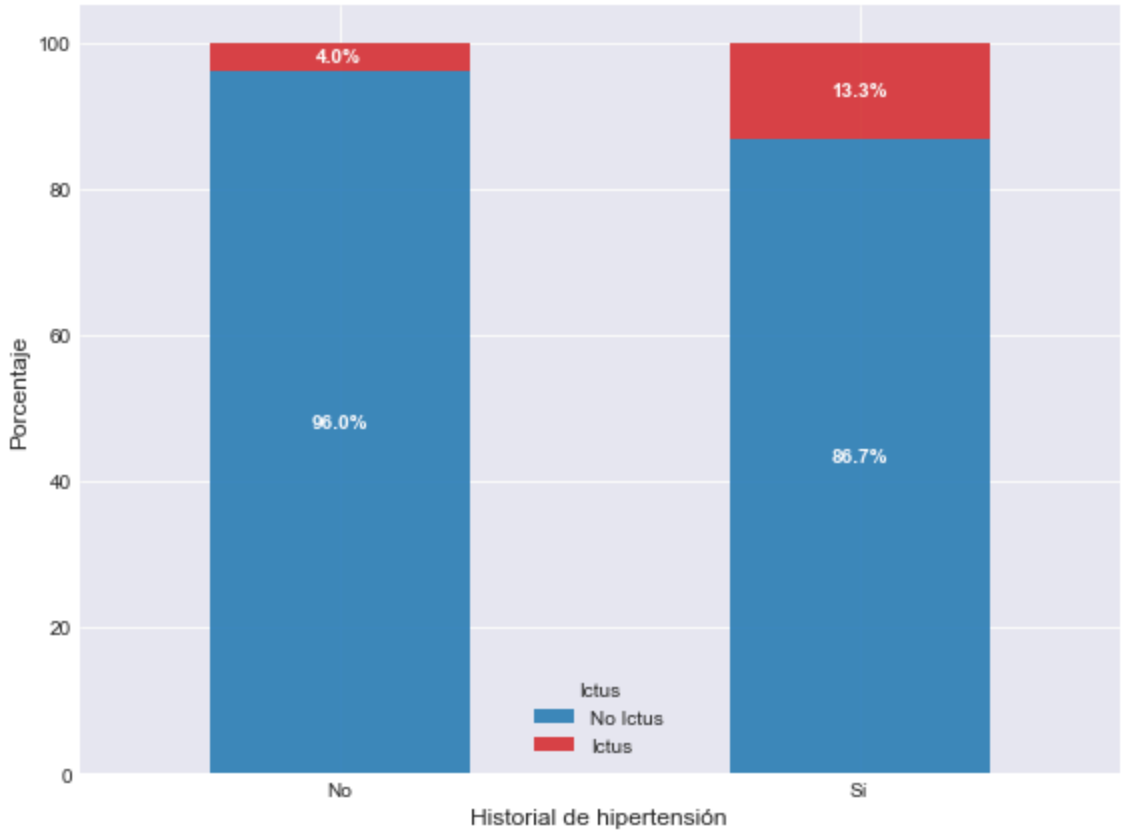

Figura 4. *Distribución porcentual del ictus según la hipertensión del paciente.*

Apreciamos pues, en la Figura 4, que el porcentaje de pacientes que han sufrido ictus se triplica en el grupo de hipertensos (Hipertensión =1) respecto del de no hipertensos, lo que parece indicar cierta relación entre hipertensión e ictus.

En la Figura 5 se muestra la proporción de pacientes que han sufrido un ictus y los que no, teniendo en cuenta si presentan antecedentes de enfermedades cardiacas o no.

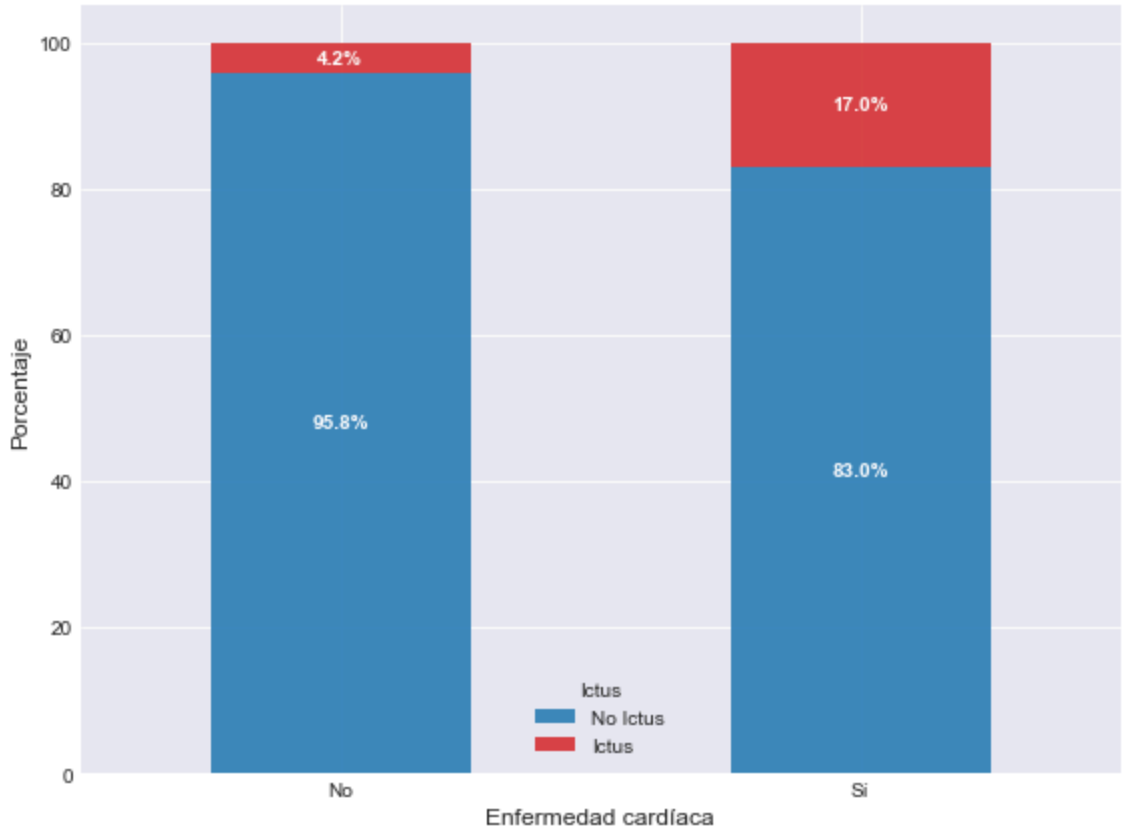

***Figura 5**: Porcentaje de ictus /no ictus para el grupo de pacientes sin antecedentes de enfermedad cardiaca (0) y con enfermedad cardiaca (1).*

Podemos llegar a una conclusión similar a la que obtuvimos sobre la relación del ictus con la variable **hipertensión,** es decir, el porcentaje de pacientes que han sufrido ictus se cuadruplica en el grupo de pacientes con antecedente de enfermedades

cardiovasculares respecto del que no los tiene, lo que parece indicar cierta relación entre enfermedades cardiacas e ictus.

En la Figura 6 se muestra la proporción de pacientes que han sufrido un ictus y los que no, según su hábito de tabaquismo

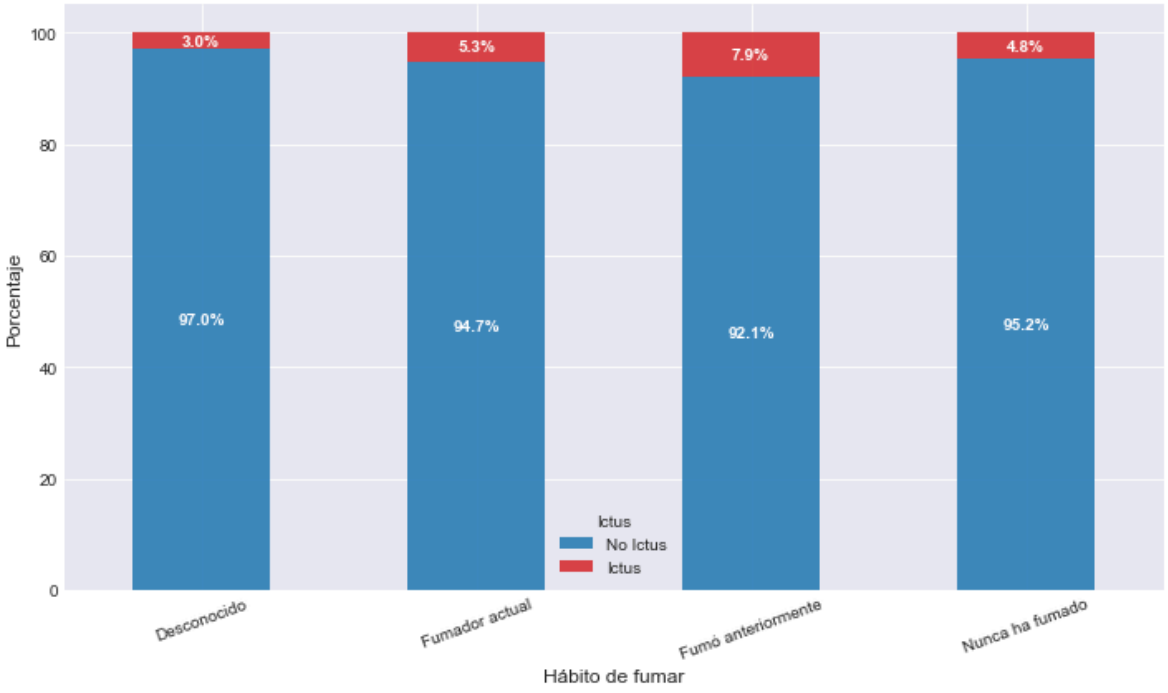

***Figura 6**: Porcentaje de ictus /no ictus dependiendo del hábito de tabaquismo del paciente.*

Apreciamos en la Figura 6, que parece haber cierta relación entre ictus y el estatus de fumador, siendo los exfumadores y fumadores los que presentan una mayor incidencia de ictus (5.3% y 7.9%, respectivamente) en comparación con quienes nunca han fumado (4.8%).

En la Figura 7 se muestra la relación entre el índice de masa corporal e ictus.

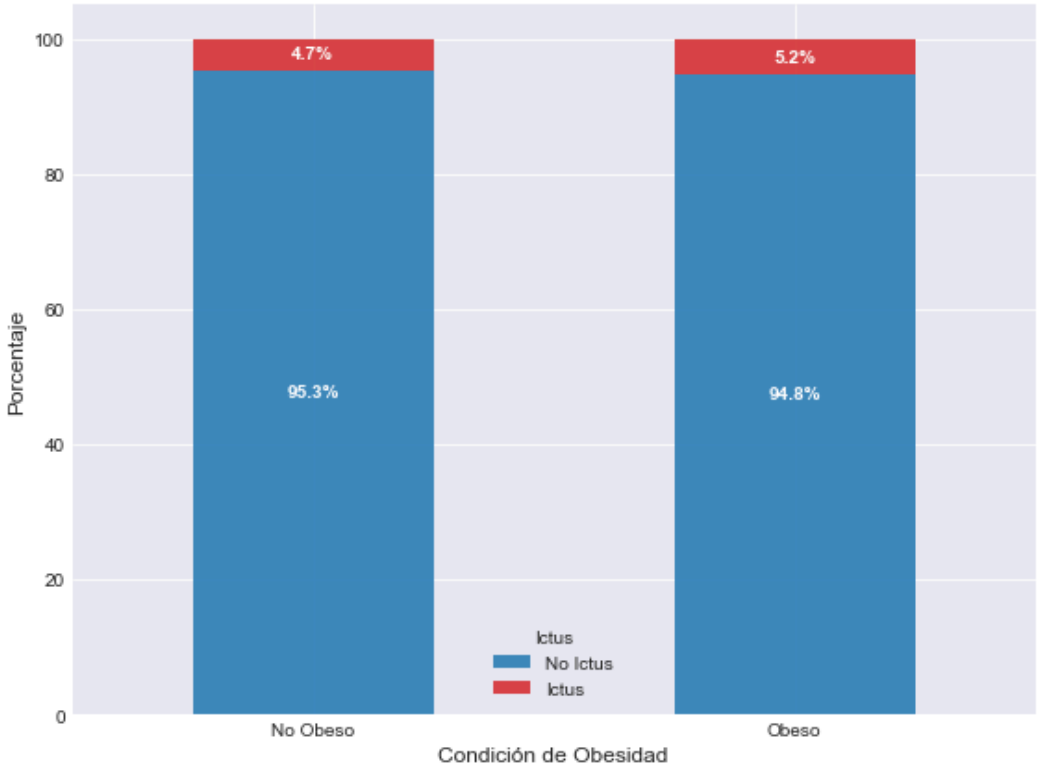

***Figura 7**: Distribución porcentual del ictus según si el paciente presenta obesidad o no.*

Apreciamos en la Figura 7, que el hecho de padecer obesidad no parece afectar a una mayor incidencia de ictus; encontramos un porcentaje de ictus similar entre los obesos (5.2%) y los no obesos (4.7%).

A continuación, indagamos la relación entre la diabetes y el ictus, diferenciando entre diabéticos y no diabéticos según el nivel de glucosa en sangre.

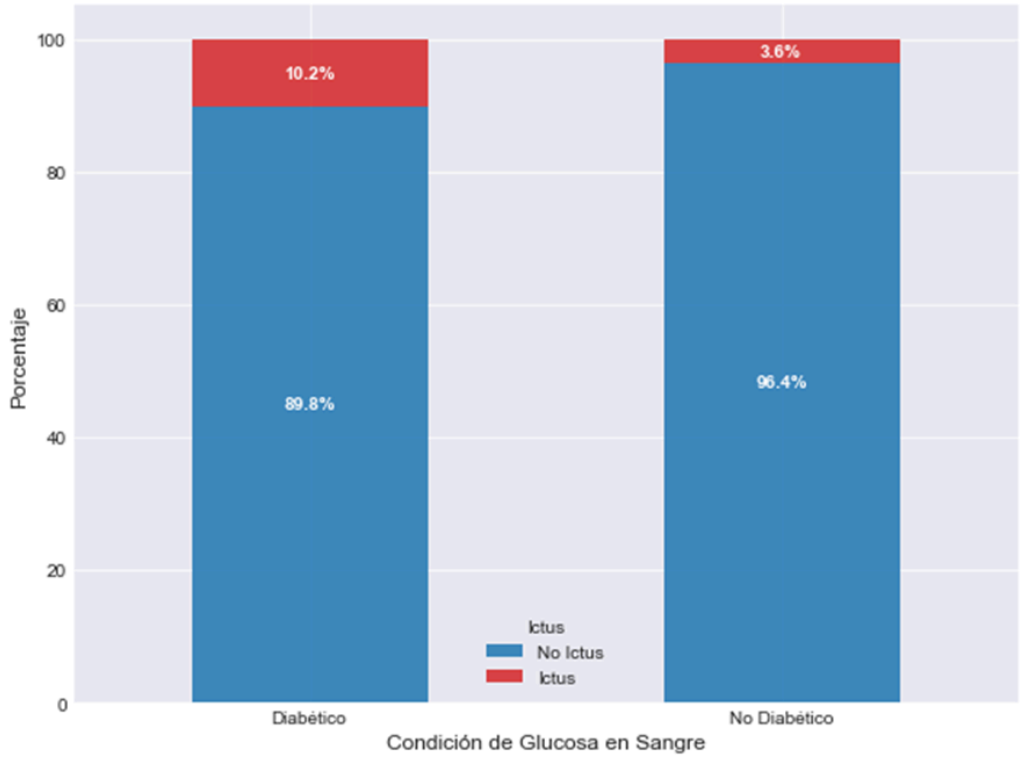

***Figura 8**: Distribución porcentual del ictus según si el paciente presenta diabetes o no.*

En la Figura 8 podemos apreciar que ser diabético va asociado a una mayor probabilidad de padecer ictus: los pacientes diabéticos presentan un 10.2% de incidencia de ictus en la muestra, y los no diabéticos un 3.6%.

Investigamos ahora la relación entre el ictus y la variable que indica si los sujetos se han casado alguna vez o no en la Figura 9.

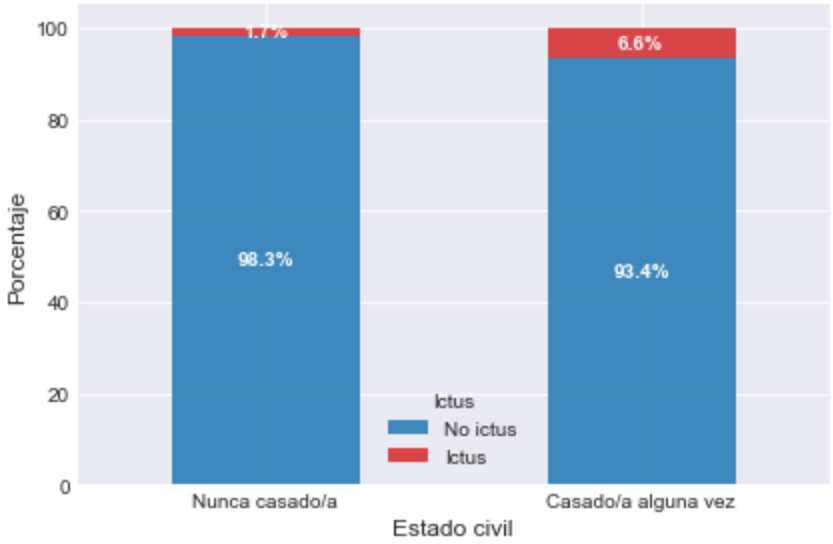

***Figura 9**: Distribución porcentual del ictus según el estado civil del paciente.*

Al parecer, las personas que sí que han estado casadas tienen más probabilidad de padecer un ictus (en torno al 6.6%) que las que no (un 1.7%). Sin embargo, hay que tener cuidado con esta relación, pues puede ser un efecto de la edad, ya que las personas que sí han estado casadas suelen ser personas mayores.

Si bien el índice de masa corporal lo vamos a utilizar a través de la variable obesidad que hemos definido. En la Figura 10, observamos la relación del índice de masa corporal con el ictus.

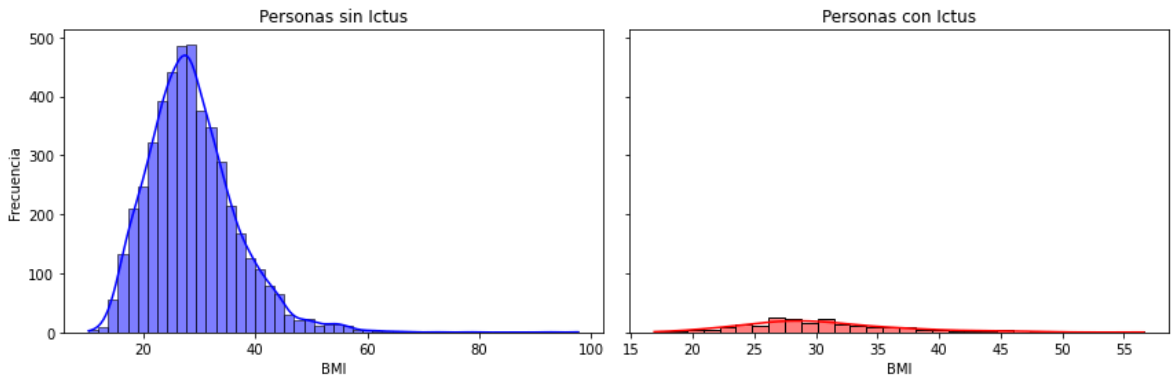

Figura 10: *Distribución del índice de masa corporal (BMI) en pacientes con y sin ictus.*

En la Figura 10 se representa la distribución del BMI en dos grupos: pacientes sin ictus (gráfico izquierdo, azul) y pacientes con ictus (gráfico derecho, rojo). Se observa que, en ambos grupos, los valores de BMI se concentran entre 20 y 35, con una forma asimétrica hacia la derecha. En el grupo sin ictus, el pico se sitúa en torno a un BMI de 28, mientras que en el grupo con ictus el pico está desplazado hacia valores algo mayores, en torno a 30.

Esto sugiere que, aunque la distribución general es parecida, los pacientes que han sufrido un ictus tienden a presentar valores de BMI ligeramente más altos.

Por último, investigamos la relación entre el ictus y la glucosa en sangre.

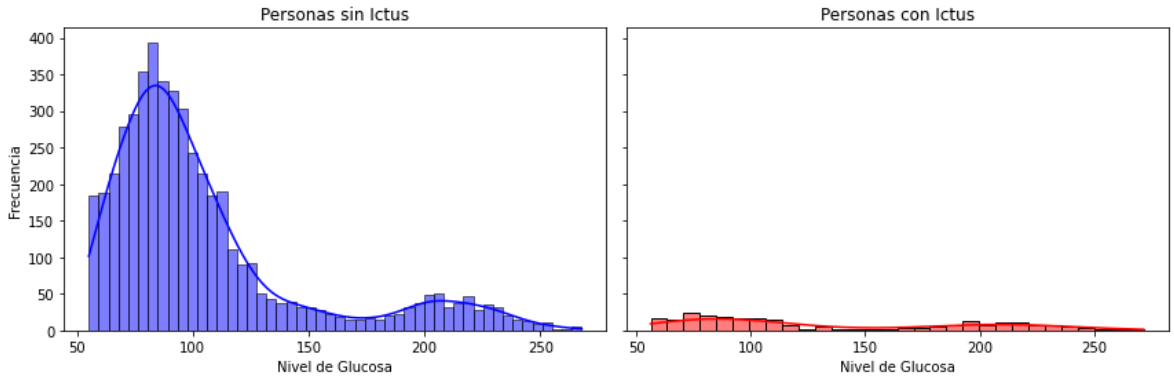

Figura 11: *Distribución del nivel de glucosa en sangre en pacientes con y sin ictus.*

La Figura 11 muestra histogramas con suavizado de densidad para comparar los niveles de glucosa entre personas que han sufrido ictus (gráfico derecho, en rojo) y las que no (gráfico izquierdo, en azul). Se aprecia que los pacientes sin ictus presentan una distribución claramente sesgada a la derecha, con un pico de frecuencia en niveles entre 90 y 110 mg/dL. Por otro lado, aunque los casos de ictus son mucho menos frecuentes, **parecen agruparse hacia niveles de glucosa más elevados**, lo que sugiere que una glucosa elevada podría ser un factor de riesgo relevante para ictus, coherente con la asociación clínica entre diabetes y eventos cerebrovasculares.

7.2 Modelización

7.2.1 Regresión logística

Se ajustaron cuatro modelos de regresión logística: uno sin regularización, y tres con penalización L1 (Lasso), L2 (Ridge) y Elastic Net. La selección de hiperparámetros se

realizó mediante validación cruzada con 10 particiones, utilizando búsqueda en malla (*Grid Search*). Los valores óptimos encontrados fueron:

- Lasso: C = 0.1
- Ridge: C = 10
- Elastic Net: C = 1

Por otro lado, respecto a la capacidad predictiva de estos modelos al clasificar a los sujetos como propensos o no a sufrir un ictus, los resultados según las métricas de clasificación y sus AICs son los siguientes:

Tabla 1A. *Métricas de clasificación y AICs de los modelos en la muestra de entrenamiento.*

Modelo	Exactitud	Precisión	Recuerdo	F1-Score	AUC	AIC
Regresión Logística	0.9513	0.0000	0.0000	0.0000	0.8448	1302.98
Ridge (L2)	0.9513	0.0000	0.0000	0.0000	0.8448	1299.40
Lasso (L1)	0.9513	0.0000	0.0000	0.0000	0.8446	1299.45

Elastic Net	0.9513	0.0000	0.0000	0.0000	0.8447	1299.21

***Tabla 1B**. Métricas de clasificación de los modelos en la muestra de prueba*

Modelo	Exactitud	Precisión	recuerdo	F1-Score	AUC
Regresión Logística	0.9511	0.0000	0.0000	0.0000	0.8397
Ridge (L2)	0.9511	0.0000	0.0000	0.0000	0.8397
Lasso (L1)	0.9511	0.0000	0.0000	0.0000	0.8394
Elastic Net	0.9511	0.0000	0.0000	0.0000	0.8396

Tal como se observa en las tablas (Tabla 1A y Tabla 1B), todos los modelos de regresión logística evaluados obtuvieron una alta exactitud tanto en la muestra train (95.13%) como en la muestra test (95.11%) y un AUC elevado (aproximadamente 0.84), lo que indica una buena capacidad de clasificación genérica, debida fundamentalmente a los pacientes sin ictus, que son mayoritarios en la muestra. Sin embargo, los valores de precisión, recuerdo y F1-score para la clase positiva (ictus)

fueron nulos, lo que refleja la incapacidad de los modelos para identificar casos de ictus.

Este resultado se debe al desequilibrio severo en la distribución de clases, que provoca que los modelos tiendan a clasificar todas las observaciones como pertenecientes a la clase mayoritaria (no ictus), es decir, no hay información suficiente para identificar correctamente a los ictus. Aunque el modelo logra una predicción global precisa, su utilidad clínica es limitada, ya que no logra detectar los eventos que precisamente se desean anticipar.

Además, en la Tabla 1A también se han incluido los valores del criterio de información de Akaike (AIC) para cada modelo, con el objetivo de valorar el equilibrio entre ajuste y complejidad. El modelo con penalización Elastic Net presentó el AIC más bajo (1299.21), esto indica que este modelo ofrece el mejor compromiso entre capacidad predictiva y parsimonia, por lo que se seleccionó como modelo final para su evaluación sobre el conjunto de prueba.

A continuación, en la Figura 12, se representaron gráficamente los coeficientes del modelo Elastic Net.

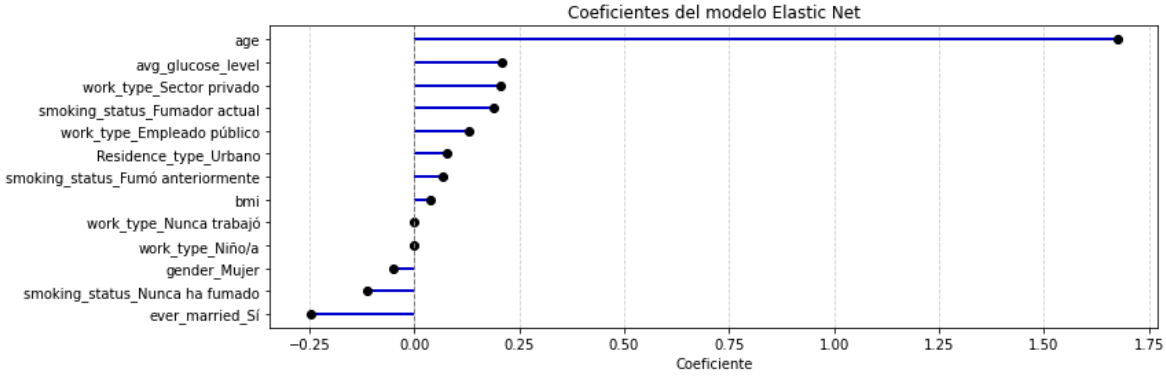

***Figura 12**: Coeficientes del modelo Elastic Net.*

La variable **edad** destaca con un coeficiente claramente mayor al resto, lo que indica que es el factor más influyente en la predicción del ictus. También se observan contribuciones positivas relevantes de **los niveles de glucosa en sangre**, el **trabajo en el sector privado** y ser **fumador actual**, lo que sugiere que estos factores incrementan la probabilidad de ictus. En cambio, variables como estar **casado** o **no haber fumado nunca** tienen efectos negativos, es decir, parecen estar asociadas a una menor probabilidad de ictus.

Por otro lado, en la Figura 13, se presenta la curva ROC del modelo sobre el conjunto de test. El área bajo la curva (AUC) refleja la capacidad discriminativa global del modelo

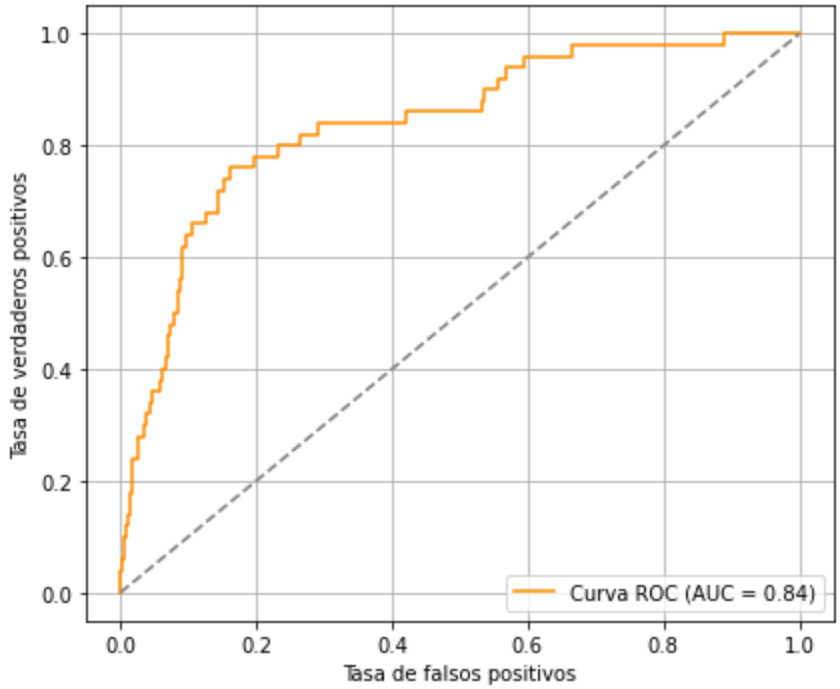

***Figura 13**: Curva ROC y área AUC para el modelo de regresión logística regularizado con elasticnet.*

La curva ROC del modelo óptimo muestra un buen desempeño en la capacidad de discriminación entre individuos con y sin ictus. El área bajo la curva (AUC) es de 0.84, lo que indica que el modelo tiene una alta capacidad para distinguir correctamente entre ambas clases. En este caso, el valor obtenido sugiere que el modelo tiene un comportamiento robusto, incluso teniendo en cuenta el desequilibrio de clases en los datos.

7.2.2 Naive Bayes

En esta sección se implementa y evalúa un modelo de clasificación basado en el algoritmo de Naïve Bayes Gaussiano. A continuación, en la Tabla 2A y en la Tabla 2B, se presentan las métricas de clasificación obtenidas:

Tabla 2A: *Resultados de clasificación del modelo Naive Bayes sobre la muestra de entrenamiento.*

Clase	Precisión	recuerdo	F1-score	Soporte
0	0.99	0.51	0.67	3888
1	0.09	0.93	0.16	199
Global				
Accuracy			0.53	
F1-score ponderado			0.65	

Tabla 2B: *Resultados de clasificación del modelo Naive Bayes sobre la muestra de prueba.*

Clase	Precisión	recuerdo	F1-score	Soporte
0	0.98	0.54	0.70	972
1	0.09	0.84	0.16	50
Global				
Accuracy		0.55		
F1-score ponderado		0.67		

El modelo muestra un desempeño desigual entre las dos clases, tanto en el conjunto de entrenamiento como en el de prueba. Para la clase 0 (no ictus), se observa una alta precisión (0.99 en entrenamiento y 0.98 en test), pero con un recuerdo moderado (0.51 y 0.54 respectivamente), lo que indica que clasifica correctamente la mayoría de los no ictus, aunque también omite una proporción relevante de ellos (casi la mitad).

Por su parte, la clase 1 (ictus) presenta una precisión muy baja (0.09 en ambos conjuntos), lo que refleja un elevado número de falsos ictus. Sin embargo, el recuerdo

es notablemente alto (0.93 en entrenamiento y 0.84 en test), lo cual sugiere que el modelo es capaz de identificar la mayoría de los casos reales de ictus.

En conjunto, aunque la exactitud global es limitada (52.9% en entrenamiento y 55.4% en test), el alto recuerdo para la clase minoritaria (ictus) es un aspecto positivo desde la perspectiva clínica, donde suele priorizarse la sensibilidad por encima de la precisión para evitar omitir posibles eventos graves. No obstante, la baja precisión del modelo implica que se generarían muchas falsas alarmas, lo que puede limitar su aplicación directa sin una etapa posterior de validación médica.

La matriz de confusión obtenida sobre el conjunto de prueba se presenta en la Figura 14:

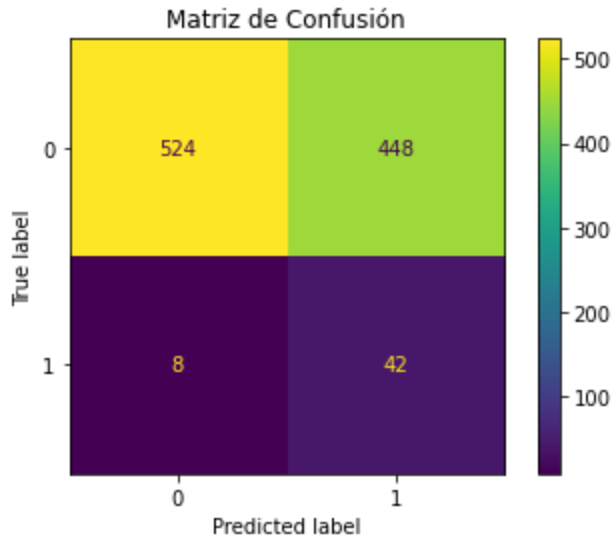

***Figura 14**: Matriz de confusión para el ajuste con el método Naive.*

El modelo clasificó correctamente **524 casos negativos** (verdaderos no ictus) y **42 casos positivos** (verdaderos ictus). Sin embargo, clasificó incorrectamente **448 negativos como positivos** (falsos positivos) y **8 positivos como negativos** (falsos negativos).

Esto concuerda con las métricas vistas anteriormente: un **recuerdo alto para la clase 1** (84%) debido a que solo 8 de los 42 casos positivos fueron omitidos, pero con una **precisión baja** porque la mayoría de las predicciones positivas fueron falsas (448 falsos positivos frente a 42 verdaderos positivos).

La matriz de confusión refuerza la idea de que el modelo, pese a su baja precisión, **detecta casi todos los casos de ictus**. Este comportamiento es deseable en sistemas de alerta temprana en medicina, donde es preferible “pasarse de precavido” (muchos falsos positivos) a omitir casos de riesgo real (falsos negativos).

Asimismo, para evaluar la capacidad del modelo de clasificación en las dos clases, se generó la **curva ROC (Receiver Operating Characteristic)**. Esta curva representa la relación entre la **tasa de verdaderos positivos (recuerdo)** y la **tasa de falsos positivos**, para diferentes umbrales de clasificación.

La siguiente figura muestra la curva ROC generada:

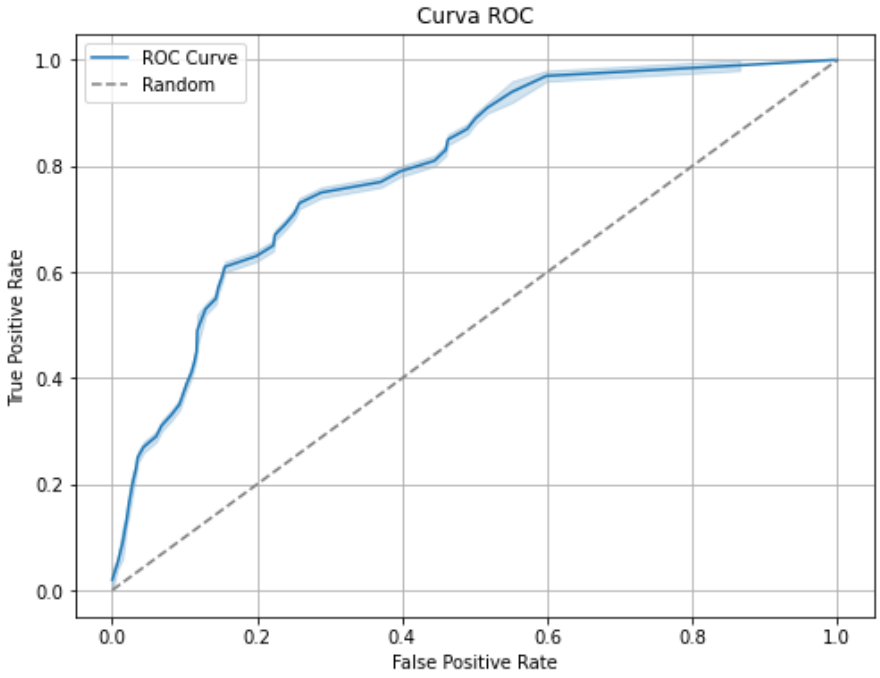

***Figura 15**: Curva ROC y área AUC del método Naive.*

El valor del área bajo la curva (AUC) obtenido es de 0.798, que indica que el modelo tiene una buena capacidad de clasificación de las clases, es decir, existe un 79.8% de probabilidad de que el modelo clasifique correctamente a un individuo elegido al azar.

Este valor refuerza lo observado previamente: aunque el modelo produce muchas falsas alarmas, tiene un buen rendimiento general clasificando correctamente los casos positivos (ictus) y muchos negativos (no ictus).

A continuación, en la Figura 16 se construye la curva de aprendizaje para valorar la oportunidad del tamaño de la muestra de entrenamiento en el ajuste.

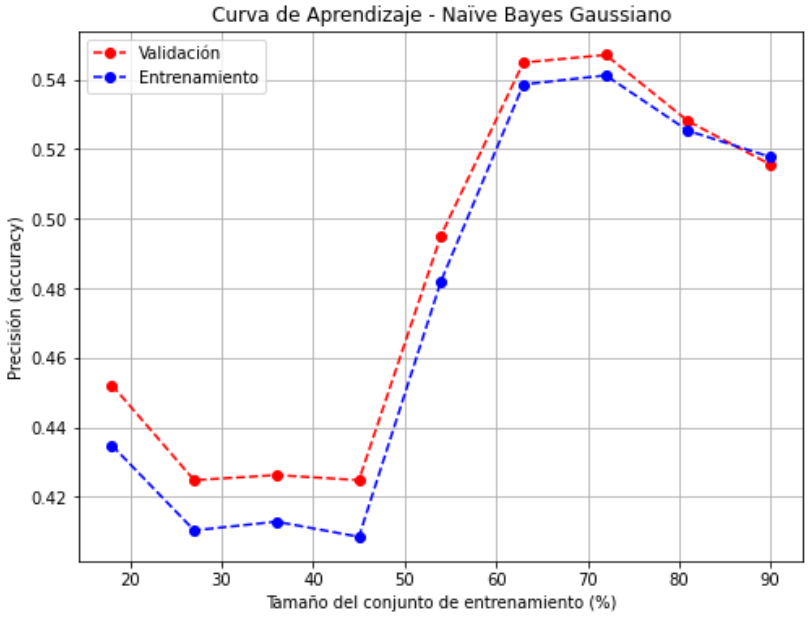

***Figura 16**: Curva de aprendizaje del método Naive.*

Se observa que tanto la **precisión en entrenamiento** como la **precisión en validación** se estabilizan en valores por encima del 50% a partir de que la muestra de entrenamiento representa al menos el 60% de los datos. Es válida pues, nuestra muestra de entrenamiento, que representa un 70% de los datos.

7.2.3. Árbol de decisión

Se ha entrenado un Árbol de Decisión sobre el conjunto de entrenamiento con los parámetros por defecto.

El modelo mostró un ajuste perfecto en la muestra de entrenamiento, obteniendo una **exactitud del 100%**, lo que sugiere un posible sobreajuste al conjunto de datos de entrenamiento.

En la muestra de test, sin embargo, el rendimiento ha sido inferior, haciendo patente el problema de sobreajuste (ver Tabla 3).

***Tabla 3**: Resultados de clasificación del modelo Árbol sobre la muestra de prueba.*

Clase / Promedio	Precisión	recuerdo	F1-score	Soporte
No ictus (0)	0.95	0.94	0.95	972
Ictus (1)	0.05	0.06	0.06	50
Exactitud	0.90			1022
Promedio ponderado	0.91	0.90	0.90	1022

Los resultados reflejan que el modelo tiende a clasificar casi todos los ejemplos como pertenecientes a la clase mayoritaria (no ictus), lo que compromete gravemente su capacidad de detección de la clase minoritaria (ictus). Esto es especialmente evidente

en el valor de **recuerdo** para ictus (0.06), indicando que el modelo solo fue capaz de identificar correctamente 3 de los 50 casos reales de esta clase.

Ante el problema de sobreajuste, se considera necesario aplicar **una poda del árbol** para reducir su complejidad y mejorar su capacidad de generalización.

Para aplicar la poda del árbol de decisión, se evaluaron 93 valores distintos del parámetro de complejidad α. Estos valores oscilaron entre 0.00000 y 0.00138. Cada modelo resultante fue evaluado sobre el conjunto de prueba, y se seleccionó como óptimo el correspondiente a α=0.00070, por ser el que ofreció la mayor exactitud.

El árbol óptimo se ajusta con los siguientes hiperparámetros:

- **Alpha óptimo:** 0.000699
- **Exactitud en entrenamiento:** 0.9513
- **Exactitud en test:** 0.9511
- **Profundidad del árbol podado:** 2
- **Número de nodos terminales:** 3

A continuación se detallan los resultados del árbol podado en los conjuntos de entrenamiento y prueba:

***Tabla 4A**: Resultados de clasificación del modelo Árbol de decisión podado sobre la muestra de entrenamiento.*

Clase / Promedio	Precisión	recuerdo	F1-score	Soporte
No ictus (0)	0.95	1.00	0.98	3888
Ictus (1)	0.00	0.00	0.00	199
Exactitud	**0.95**			4087
Promedio ponderado	0.90	0.95	0.93	

***Tabla 4B**: Resultados de clasificación del modelo Árbol de decisión podado sobre la muestra de prueba.*

Clase	Precisión	recuerdo	F1-score	Soporte
No ictus (0)	0.95	1.00	0.97	972

Ictus (1)	0.00	0.00	0.00	50
Exactitud			**0.95**	1022
Promedio ponderado	0.90	0.95	0.93	1022

A pesar de que la poda ha permitido obtener un modelo mucho más simple y con menor riesgo de sobreajuste (reduciendo la profundidad del árbol y el número de nodos), el comportamiento predictivo sobre los casos de ictus es completamente deficiente. El modelo no clasifica ningún caso de ictus correctamente, clasificando todos los registros como no ictus.

Este comportamiento indica que el árbol de decisión no es adecuado para esta base de datos, ya que prioriza la exactitud global clasificando correctamente la mayoría de no ictus, pero **no consigue clasificar bien ningún caso de ictus**, que en este contexto son probablemente los casos más importantes de detectar. Esto lo convierte en un modelo inapropiado para tareas donde el objetivo es precisamente detectar de forma temprana y precisa los eventos de ictus.

Por último, en la Figura 17 se muestra la importancia de las variables en el árbol de decisión podado.

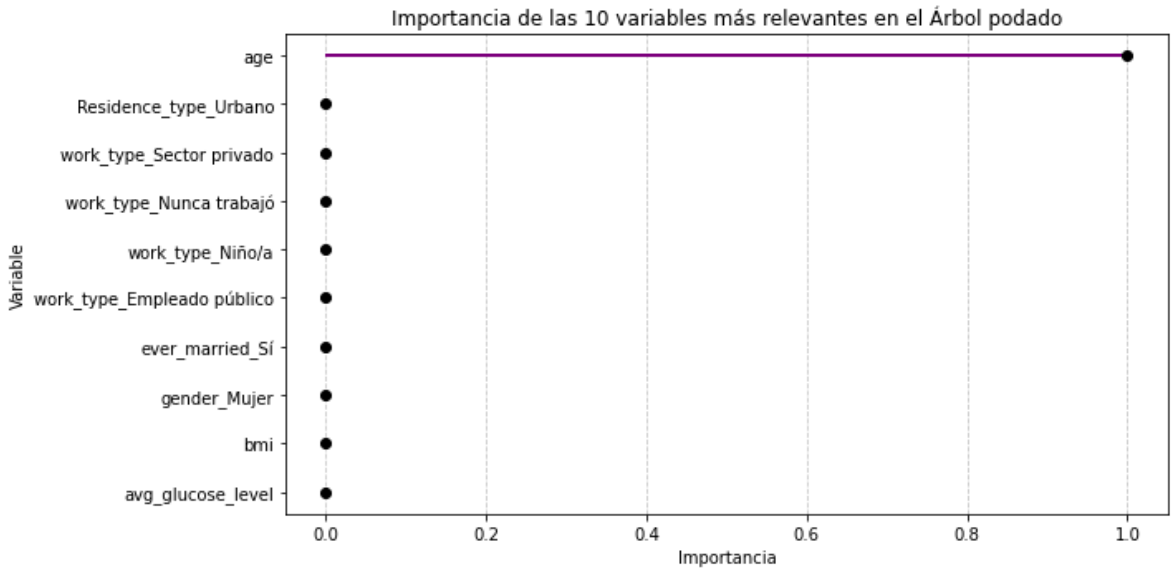

Figura 17. *Importancia de las variables en el árbol de decisión podado.*

El resultado muestra que únicamente la variable age (edad) presenta una importancia distinta de cero, siendo la única utilizada por el modelo para realizar particiones. Esto se debe a que el árbol podado tiene una profundidad máxima de 2, lo que limita considerablemente su capacidad para incorporar múltiples variables. Como consecuencia, el modelo elige únicamente aquella variable que proporciona la mayor ganancia de información en el primer nivel, en este caso la edad. Este resultado refleja un modelo extremadamente simple, insuficiente para capturar relaciones complejas en los datos.

7.2.4 Random forest

Se realizó una búsqueda en malla (*Grid Search*) para ajustar el modelo de Random Forest, explorando combinaciones de los siguientes hiperparámetros:

- n_estimators: número de árboles en el bosque (se fijó en 150).
- max_features: número de predictores considerados en cada división del árbol (valores evaluados: 1 y 2)
- max_depth: profundidad máxima de los árboles (valores evaluados: 2, 3, 5, 10 y 20)
- criterion: criterio de impureza (valores evaluados: 'gini' y 'entropy')

En total, se evaluaron 10 combinaciones distintas, utilizando como métrica de selección la exactitud *out-of-bag* (OOB), que permite estimar el rendimiento del modelo sin necesidad de un conjunto de validación externa. La combinación óptima hallada fue:

- Criterio: **Gini**
- Profundidad máxima: **20**
- Variables por división: **1**
- Número de árboles: **150**

A continuación se presentan sus métricas de evaluación:

***Tabla 5**: Resultados de clasificación del modelo Random Forest sobre la muestra de prueba.*

Clase	Precisión	recuerdo	F1-score	Soporte
No ictus	0.95	1.00	0.97	972
Ictus	0.00	0.00	0.00	50
Promedios globales				
Accuracy		0.95		
F1-score ponderado		0.93		

Matriz de confusión:

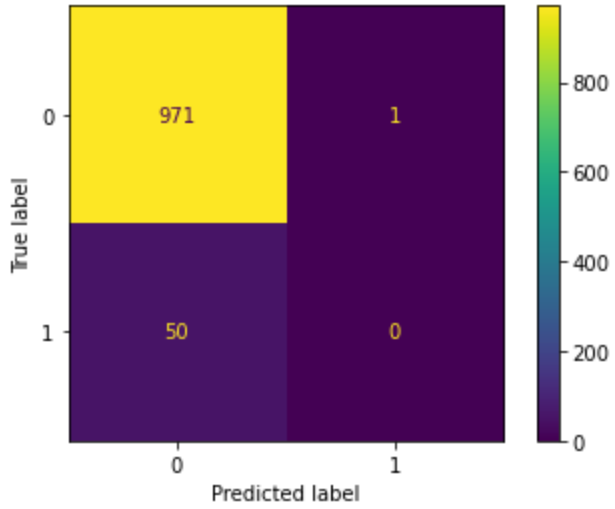

Figura 18: *Matriz de confusión del método Random Forest.*

El modelo clasificó correctamente 971 de los 972 pacientes que no sufrieron ictus, pero no fue capaz de detectar ninguno de los 50 casos positivos (ictus).

Aunque el modelo Random Forest presenta una alta precisión global debido a su correcto desempeño en la clase mayoritaria (no ictus), su incapacidad total para identificar casos de ictus lo convierte en un modelo clínicamente inadecuado para esta tarea. Esta deficiencia se explica por el fuerte desequilibrio de clases presente en los datos, en los que los pacientes con ictus representan menos del 5% del total.

Por último, en la Figura 19 se muestra la importancia de las variables en el módelo óptimo del Random Forest. .

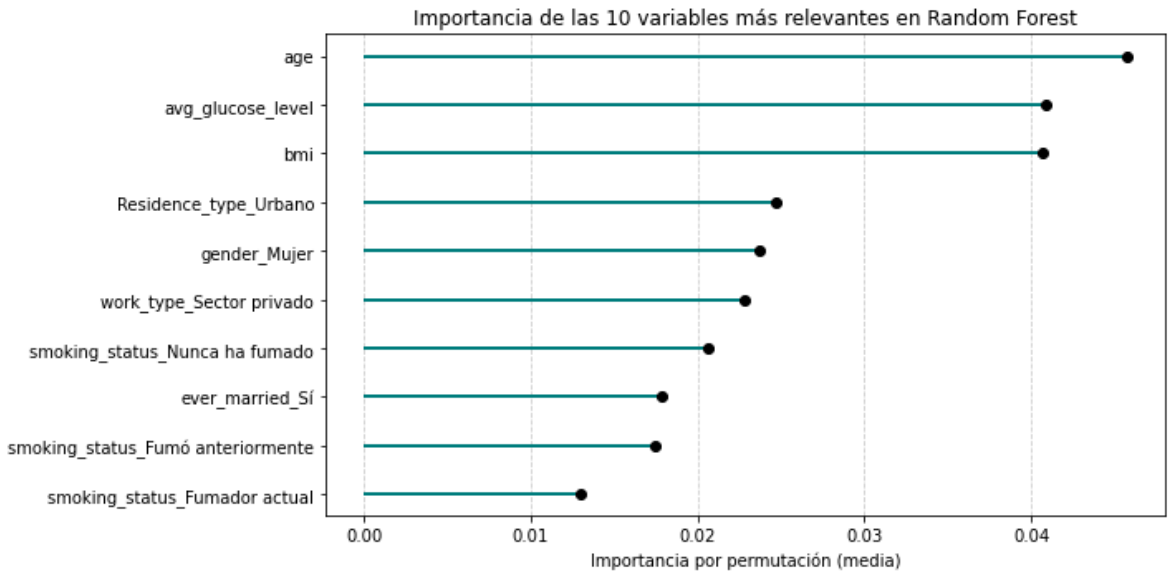

Figura 19. *Importancia de las variables en el modelo óptimo del modelo Random Forest.*

En la Figura 19, se observa la importancia de las 10 variables más relevantes según el modelo de Random Forest, evaluada mediante la técnica de permutación. Destacan como predictores más influyentes la edad, el nivel medio de glucosa y el índice de masa corporal (BMI), lo cual resulta coherente con la literatura clínica sobre factores de riesgo del ictus. A diferencia del árbol de decisión podado (Figura 19), aquí se observa que múltiples variables contribuyen al modelo, reflejando la mayor capacidad del Random Forest para captar relaciones complejas y utilizar información conjunta de distintas características.

8. Conclusiones

El objetivo principal de este trabajo fue desarrollar modelos predictivos capaces de estimar la probabilidad de que un paciente sufra un ictus, a partir de variables clínicas y demográficas. Para ello, se utilizó el conjunto de datos *Stroke Prediction Dataset* y se aplicaron técnicas de aprendizaje automático supervisado, incluyendo regresión logística (con y sin regularización), Naïve Bayes, árbol de decisión y Random Forest.

Los resultados obtenidos permiten extraer varias conclusiones relevantes:

- **Desequilibrio de clases**: La base de datos presentaba un fuerte desequilibrio en los tamaños muestrales de pacientes con ictus (clase minoritaria) y sin ictus (clase mayoritaria). Este desequilibrio afecta la capacidad de los modelos para clasificar correctamente los casos de ictus, generando modelos con alta exactitud global pero muy baja sensibilidad.

- **Modelos tradicionales como la regresión logística** obtuvieron un rendimiento global adecuado (AUC ≈ 0.84), pero fueron incapaces de detectar casos de ictus (recuerdo = 0). Esto limita su utilidad en contextos clínicos donde la detección de eventos positivos es prioritaria.

- El modelo **Naïve Bayes**, aunque mostró menor exactitud general, fue el único capaz de alcanzar una **alta sensibilidad (84%)** en la clasificación de ictus, a costa de una precisión muy baja. Esto sugiere que, en entornos médicos donde se prioriza la detección temprana, podría utilizarse como modelo de prueba

inicial, complementado por una segunda fase de validación clínica.

- **Árbol de decisión y Random Forest** mostraron una fuerte tendencia a predecir exclusivamente la clase mayoritaria (no ictus), clasificando erróneamente todos los casos de ictus. A pesar de su capacidad para manejar relaciones no lineales, estos modelos resultaron ineficaces con estos datos.

- A nivel de **importancia de variables**, la edad, los niveles de glucosa en sangre y el índice de masa corporal (BMI) surgieron consistentemente como los predictores más relevantes en el modelo de bosques, y en el modelo de regresión logística regularizado (Elastic Net) destacaron también la edad como el factor más influyente, seguido por el nivel de glucosa, el hecho de ser fumador actual y trabajar en el sector privado, todos ellos con coeficientes positivos. En sentido contrario, variables como no haber fumado nunca o haber estado casado mostraron coeficientes negativos, lo que sugiere una posible asociación con menor riesgo de ictus. También se observaron asociaciones destacables con el tabaquismo, la hipertensión y antecedentes de enfermedad cardíaca, lo que concuerda con la literatura médica existente.

En conclusión, aunque los modelos desarrollados presentan limitaciones importantes, especialmente debido al desequilibrio de clases, este estudio demuestra que las técnicas de aprendizaje automático pueden ser útiles como herramienta complementaria en la identificación de factores de riesgo del ictus. Para mejorar su rendimiento y aplicabilidad en entornos clínicos reales, se recomienda ampliar el

tamaño de los pacientes que han padecido ictus.

Este trabajo sienta las bases para estudios futuros en los que se profundice en la personalización del riesgo y la toma de decisiones clínicas apoyadas por modelos de inteligencia artificial.

References

Breiman, L. (2001). *Random forests.* Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC. https://link.springer.com/book/10.1007/978-1-4899-7983-7

Cerebrovascular accident (stroke). (2023, mayo 9). Organización Mundial de la Salud.

https://www.who.int/news-room/fact-sheets/detail/cerebrovascular-accident-(stroke)

Ferretti, J. (s.f.). (n.d.). *Causes of Stroke – Log Regr, Partial Dependence & SHAP [Notebook].* Kaggle. https://www.kaggle.com/code/jacopoferretti/causes-of-stroke-log-regr-partial-dependence-shap

Hankey, G. J. (2017). *Stroke.* The Lancet, 389(10069), 641–654. https://doi.org/10.1016/S0140-6736(16)30962-X

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). *Array programming with NumPy. Nature, 585(7825), 357–362*. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment.* Computing in Science & Engineering, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Índice de masa corporal. (2025). MedlinePlus. Retrieved April 7, 2025, from https://medlineplus.gov/spanish/ency/article/007196.htm

Kse, A. (s.f.). (n.d.). *Stroke Data EDA [Notebook]*. Kaggle. https://www.kaggle.com/code/abdullahkse/stroke-data-eda

Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions.* Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

McKinney, W. (2010). *Data structures for statistical computing in Python.* Proceedings of the 9th Python in Science Conference, 51–56. https://doi.org/10.25080/Majora-92bf1922-00a

MedlinePlus. (2006). *Global variation in stroke burden and mortality: estimates from monitoring, surveillance, and modelling.* Johnston, S. C., Mendis, S., & Mathers, C. D. https://doi.org/10.1016/S1474-4422(06)70400-3

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable.* https://christophm.github.io/interpretable-ml-book/

Nivel de glucosa en sangre. (2024, 28 02). MedinePlus. https://medlineplus.gov/spanish/ency/patientinstructions/000086.htm

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.* https://scikit-learn.org/

Quinlan, J. R. (1986). *Induction of decision trees.* Machine Learning, 1(1), 81–106. https://doi.org/10.1007/BF00116251

Raschka, S. (2020). *Machine Learning with Python: Learn how to build and evaluate machine learning models using scikit-learn.* https://sebastianraschka.com/blog/2020/learning-curves.html

Soriano, F. (s.f.). (n.d.). *Stroke Prediction Dataset [Dataset].* Kaggle. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Waskom, M. (2021). *seaborn: Statistical data visualization.* Journal of Open Source Software, 6(60), 3021. https://doi.org/10.21105/joss.03021

Códigos usados de los modelos en python:

- Árbol de decisión
- Regresión logística
- Naïve Bayes
- Random Forest

Conforme a la última edición de las normas APA (7ª Edición en 2025).