Centro Europeo de Másteres y Posgrados (CEMP)

Máster en Bioestadística y Bioinformática

Trabajo Fin de Máster

Predicción de la ocurrencia de infarto cerebral

Directora: Sara López

Curso académico 2024–2025

January 5, 2026

Índice

Índice de tablas

Índice de figuras

Abreviaturas empleadas

BMI: Índice de masa corporal (Body Mass Index)
EDA: Análisis exploratorio de datos (Exploratory Data Analysis)
IMC: Índice de masa corporal
ML: Machine Learning
OR: Odds Ratio
ROC: Receiver Operating Characteristic
RNA-seq: Secuenciación de ARN
TFM: Trabajo Fin de Máster
WHO: World Health Organization

1 Resumen

El infarto cerebral constituye una de las principales causas de mortalidad y discapacidad a nivel mundial y representa un importante problema de salud pública debido a su elevada incidencia y a las secuelas funcionales que genera. La identificación temprana de individuos con alto riesgo resulta fundamental para el diseño de estrategias preventivas eficaces y para la optimización de los recursos sanitarios.

El objetivo de este Trabajo Fin de Máster fue analizar un conjunto de datos clínicos y demográficos con el fin de identificar factores asociados a la ocurrencia de infarto cerebral y desarrollar modelos predictivos capaces de estimar la probabilidad de aparición de la enfermedad. Para ello, se llevó a cabo un análisis estadístico que incluyó estadística descriptiva, contrastes de hipótesis y análisis de correlación, complementado con el desarrollo de un modelo de regresión logística como herramienta predictiva.

El conjunto de datos analizado incluyó variables demográficas, clínicas y metabólicas, tales como edad, sexo, hipertensión arterial, niveles de glucosa en sangre, índice de masa corporal y hábitos de tabaquismo. Los resultados mostraron diferencias estadísticamente significativas entre individuos con y sin infarto cerebral para variables como la edad, la hipertensión arterial y la glucosa media. En el modelo de regresión logística, la edad y la hipertensión se identificaron como los principales factores de riesgo asociados, mientras que otras variables no alcanzaron significación estadística tras el ajuste multivariable.

La evaluación del rendimiento del modelo evidenció un comportamiento influido por el desbalanceo del conjunto de datos, observándose una elevada especificidad con el umbral de clasificación estándar y un aumento de la sensibilidad al emplear un umbral alternativo. Estos resultados ponen de manifiesto la importancia de adaptar los modelos predictivos al contexto clínico de aplicación.

En conjunto, este trabajo demuestra el potencial del análisis estadístico y de los modelos predictivos basados en datos clínicos como herramientas de apoyo para la identificación de factores de riesgo y la predicción del infarto cerebral, contribuyendo a una medicina más preventiva y orientada al riesgo.

2 Abstract

Stroke is one of the leading causes of mortality and long-term disability worldwide and represents a major public health challenge due to its high incidence and associated functional sequelae. Early identification of individuals at high risk is essential for the implementation of effective preventive strategies and the optimization of healthcare resources.

The aim of this Master's Thesis was to analyze a clinical and demographic dataset in order to identify factors associated with the occurrence of stroke and to develop predictive models capable of estimating the probability of disease onset. To this end, a comprehensive statistical analysis was conducted, including descriptive statistics, hypothesis testing and correlation analysis, followed by the development of a logistic regression model as a predictive tool.

The dataset included demographic, clinical and metabolic variables such as age, sex, hypertension, average blood glucose levels, body mass index and smoking status. The results revealed statistically significant differences between individuals with and without stroke, particularly with respect to age, hypertension and glucose levels. In the logistic regression model, age and hypertension emerged as the main risk factors associated with stroke occurrence, while other variables did not reach statistical significance after multivariable adjustment.

Model performance evaluation showed an influence of class imbalance, with high specificity observed at the standard classification threshold and improved sensitivity when an alternative threshold was applied. These findings highlight the importance of adapting predictive models to their intended clinical context.

Overall, this study demonstrates the potential of statistical analysis and data-driven predictive models as support tools for risk assessment and prevention in cerebrovascular disease, contributing to a more proactive and risk-oriented approach to clinical decision-making.

3 Agradecimientos

La realización de este Trabajo Fin de Máster ha supuesto un reto importante, tanto a nivel profesional como personal, y no habría sido posible sin el apoyo de las personas que me han acompañado durante este proceso.

Quiero agradecer de manera muy especial a mis hijos, Carmelo y Raúl, por su paciencia y comprensión durante todo este tiempo. A pesar de los momentos en los que el trabajo y el estudio me han robado horas con ellos, su cariño y su manera de estar siempre presentes han sido un motor fundamental para seguir adelante. Este trabajo es también para ellos, con la esperanza de que les sirva como ejemplo de que el esfuerzo, la constancia y el compromiso con los propios objetivos merecen la pena.

Mi agradecimiento más sincero a mi marido,Carmelo, por su apoyo incondicional, por su comprensión y por haber estado a mi lado en los momentos de mayor carga y cansancio. Su acompañamiento ha sido clave para poder seguir adelante incluso cuando parecía que no llegaba a todo.

Quiero agradecer de forma muy especial a mi madre, por su apoyo constante y por sus palabras en los momentos más difíciles. Cuando me he sentido desbordada, cansada o con la sensación de no estar llegando a todo, siempre ha sabido recordarme que podía hacerlo y que estaba dando lo mejor de mí. Su confianza y su forma de sostenerme emocionalmente han sido fundamentales para no rendirme.

Por último, agradezco al Centro Europeo de Másteres y Posgrados y al profesorado del Máster en Bioestadística y Bioinformática la formación recibida, que ha permitido adquirir los conocimientos necesarios para el desarrollo de este trabajo.

4 Introducción

El infarto cerebral, también conocido como ictus, constituye una de las principales causas de mortalidad y discapacidad a nivel mundial. Se trata de una patología de origen multifactorial caracterizada por la interrupción del flujo sanguíneo cerebral, lo que provoca daño neuronal y alteraciones neurológicas de gravedad variable. En función de su etiología, el ictus puede clasificarse principalmente en isquémico, causado por la obstrucción de un vaso sanguíneo, o hemorrágico, originado por la rotura de dicho vaso. El infarto cerebral isquémico representa aproximadamente el 80–85 % de los casos y supone un importante problema de salud pública debido a su elevada incidencia y a las secuelas funcionales que con frecuencia genera (Feigin et al., 2017; Powers et al., 2018; World Health Organization, 2023).

Diversos factores de riesgo han sido asociados al desarrollo de infarto cerebral, entre los que se incluyen la edad avanzada, la hipertensión arterial, la diabetes mellitus, las enfermedades cardiovasculares, el tabaquismo y determinadas condiciones socioeconómicas. La identificación temprana de individuos con alto riesgo resulta fundamental para la implementación de estrategias preventivas y para la reducción de la carga clínica y social asociada a esta enfermedad. En este contexto, el análisis de datos clínicos y demográficos se ha convertido en una herramienta clave para mejorar la comprensión de los mecanismos subyacentes al infarto cerebral y para apoyar la toma de decisiones clínicas (Goldstein et al., 2011; O'Donnell et al., 2016).

En los últimos años, el avance de las técnicas estadísticas y de aprendizaje automático ha permitido el desarrollo de modelos predictivos capaces de identificar patrones complejos en grandes conjuntos de datos biomédicos. Estas aproximaciones han demostrado ser especialmente útiles en el ámbito de la medicina predictiva, donde el objetivo no solo es describir asociaciones entre variables, sino también estimar la probabilidad de que un individuo desarrolle una determinada patología. En el caso del infarto cerebral, numerosos estudios han explorado el uso de modelos de regresión logística, árboles de decisión, random forest y máquinas de soporte vectorial para predecir su ocurrencia a partir de factores de riesgo clínicos y demográficos (Hosmer et al., 2013; Steyerberg, 2019).En este contexto, trabajos recientes han mostrado que los enfoques basados en machine learning pueden mejorar la identificación de pacientes con alto riesgo de infarto cerebral y aportar valor añadido en la predicción de resultados clínicos frente a métodos tradicionales (Heo et al., 2019).

No obstante, a pesar del creciente número de investigaciones en este ámbito, persisten desafíos relacionados con la interpretación de los resultados, la selección de variables relevantes y la comparación objetiva del rendimiento de distintos modelos predictivos. Además, resulta imprescindible complementar los enfoques de aprendizaje automático con análisis estadísticos clásicos, como los contrastes de hipótesis y el estudio de correlaciones, que permitan evaluar de manera rigurosa la

relación entre los factores de riesgo y la enfermedad (Hosmer et al., 2013; Steyerberg, 2019).

En este trabajo se aborda el análisis de un conjunto de datos clínicos relacionados con el infarto cerebral con el objetivo de desarrollar modelos predictivos reproducibles y evaluar su capacidad para estimar la probabilidad de ocurrencia de la enfermedad. Mediante la combinación de análisis estadístico y técnicas de aprendizaje automático, se pretende aportar una visión integral que contribuya a una mejor comprensión de los factores asociados al infarto cerebral y al potencial de los modelos predictivos en el ámbito de la salud (Steyerberg, 2019).

4.1 1.1. Infarto cerebral: definición y clasificación

El infarto cerebral, comúnmente denominado ictus isquémico, se produce como consecuencia de la interrupción del flujo sanguíneo hacia una región del cerebro, lo que origina un déficit de oxígeno y nutrientes y conduce a la muerte celular neuronal. Este proceso desencadena una cascada de acontecimientos fisiopatológicos que incluyen excitotoxicidad, inflamación y daño oxidativo, responsables en última instancia de las secuelas neurológicas observadas tras el evento agudo.

Desde el punto de vista clínico y etiológico, el ictus se clasifica de manera general en dos grandes categorías: ictus isquémico e ictus hemorrágico. El ictus isquémico, que representa aproximadamente el 80–85 % de los casos, se origina por la oclusión de una arteria cerebral, ya sea por un trombo formado localmente o por un émbolo procedente de otra localización del sistema vascular. Por el contrario, el ictus hemorrágico se debe a la rotura de un vaso sanguíneo intracraneal, provocando extravasación de sangre en el parénquima cerebral o en los espacios meníngeos.

Dentro del ictus isquémico, existen diversas clasificaciones en función del mecanismo subyacente. Una de las más utilizadas en la práctica clínica es la clasificación TOAST, que distingue entre ictus aterotrombótico, cardioembólico, lacunar, de causa determinada y de causa indeterminada. Esta clasificación resulta especialmente relevante, ya que cada subtipo se asocia a perfiles de riesgo, pronóstico y estrategias terapéuticas diferentes.

El ictus aterotrombótico se relaciona principalmente con la presencia de placas de ateroma en grandes arterias, como la carótida interna, mientras que el ictus cardioembólico suele estar asociado a patologías cardíacas, entre las que destaca la fibrilación auricular. El ictus lacunar, por su parte, afecta a pequeñas arterias perforantes y se vincula estrechamente con la hipertensión arterial crónica y la diabetes mellitus. Esta heterogeneidad etiológica subraya el carácter multifactorial del infarto cerebral y justifica la necesidad de modelos predictivos que integren múltiples variables clínicas y demográficas.

4.2 Epidemiología y carga sanitaria del infarto cerebral

El infarto cerebral constituye uno de los principales problemas de salud pública a nivel mundial debido a su elevada incidencia, mortalidad y al importante impacto funcional que genera en la población afectada. Según la Organización Mundial de la Salud, el ictus se sitúa entre las primeras causas de muerte y representa una de las principales causas de discapacidad adquirida en adultos, especialmente en países desarrollados.

A nivel global, se estima que cada año se producen más de 12 millones de nuevos casos de ictus, de los cuales aproximadamente 6,5 millones resultan mortales. Aunque en las últimas décadas se ha observado una reducción de la mortalidad asociada al ictus en algunos países de altos ingresos, esta disminución se ha visto acompañada de un aumento en la prevalencia de personas que sobreviven con secuelas neurológicas, lo que incrementa la carga asistencial y socioeconómica de la enfermedad.

En el contexto europeo, el infarto cerebral continúa siendo una de las principales causas de mortalidad cardiovascular, solo por detrás de la cardiopatía isquémica. En España, el ictus representa la primera causa de muerte en mujeres y la segunda en hombres, así como la principal causa de discapacidad grave en adultos. Se estima que cada año se producen en torno a 110.000 nuevos casos, con una incidencia que aumenta de forma marcada con la edad.

El envejecimiento progresivo de la población constituye uno de los factores clave que explican el aumento esperado del número absoluto de casos de infarto cerebral en los próximos años. A este fenómeno se suma la elevada prevalencia de factores de riesgo modificables, como la hipertensión arterial, la diabetes mellitus, la obesidad y el sedentarismo, que contribuyen de manera significativa al desarrollo de la enfermedad.

Desde el punto de vista económico, el impacto del ictus es considerable. Los costes directos incluyen la atención hospitalaria en fase aguda, los tratamientos farmacológicos y las intervenciones de rehabilitación, mientras que los costes indirectos abarcan la pérdida de productividad laboral, la dependencia funcional y la necesidad de cuidados de larga duración. Este elevado coste sanitario y social refuerza la importancia de desarrollar estrategias de prevención y de identificación temprana de individuos con alto riesgo de infarto cerebral.

En este contexto, el uso de herramientas predictivas basadas en datos clínicos y demográficos adquiere una relevancia creciente, ya que permite orientar medidas preventivas, optimizar recursos sanitarios y mejorar el pronóstico de los pacientes mediante intervenciones precoces.

4.3 Factores de riesgo del infarto cerebral

El desarrollo del infarto cerebral está estrechamente relacionado con la presencia de múltiples factores de riesgo, tanto modificables como no modificables. La identificación y el control de estos factores constituye uno de los pilares fundamentales en la prevención primaria y secundaria de la enfermedad. En las últimas décadas, numerosos estudios epidemiológicos han permitido establecer una relación clara entre determinados factores clínicos y demográficos y el riesgo de sufrir un ictus.

4.3.1 Edad y sexo

La edad es uno de los factores de riesgo no modificables más relevantes en el infarto cerebral. La incidencia del ictus aumenta de forma exponencial a partir de los 55–60 años, duplicándose aproximadamente cada década. Este incremento se asocia a cambios fisiológicos propios del envejecimiento, como el deterioro vascular, la mayor rigidez arterial y la acumulación de comorbilidades cardiovasculares.

En cuanto al sexo, se ha observado que los hombres presentan una mayor incidencia de infarto cerebral a edades más tempranas, mientras que en las mujeres el riesgo aumenta de forma significativa a partir de la menopausia. No obstante, debido a una mayor esperanza de vida, las mujeres presentan un mayor número absoluto de casos y una mayor mortalidad asociada al ictus. Estas diferencias justifican la inclusión del sexo como variable relevante en los modelos predictivos.

4.3.2 Hipertensión arterial

La hipertensión arterial es el principal factor de riesgo modificable del infarto cerebral y se asocia tanto a ictus isquémicos como hemorrágicos. La elevación mantenida de la presión arterial provoca daño en la pared vascular, favoreciendo la aterosclerosis, la formación de trombos y la rotura de pequeños vasos cerebrales.

Diversos estudios han demostrado que el control adecuado de la presión arterial reduce de forma significativa el riesgo de infarto cerebral, incluso en personas de edad avanzada. Por este motivo, la hipertensión arterial constituye una variable clave en cualquier análisis orientado a la predicción del riesgo de ictus.

4.3.3 Diabetes mellitus y niveles de glucosa

La diabetes mellitus se asocia a un aumento del riesgo de infarto cerebral debido a los efectos de la hiperglucemia crónica sobre el endotelio vascular. La alteración del metabolismo de la glucosa

favorece la inflamación, el estrés oxidativo y el desarrollo acelerado de aterosclerosis.

Además, incluso en individuos sin diagnóstico previo de diabetes, niveles elevados de glucosa en sangre se han relacionado con un peor pronóstico y una mayor incidencia de eventos cerebrovasculares. Por ello, los niveles medios de glucosa constituyen un marcador clínico de gran interés en estudios epidemiológicos y modelos predictivos.

4.3.4 Obesidad e índice de masa corporal

La obesidad, habitualmente evaluada mediante el índice de masa corporal (IMC), se considera un factor de riesgo indirecto del infarto cerebral. El exceso de tejido adiposo se asocia a un mayor riesgo de desarrollar hipertensión, diabetes mellitus, dislipemia y enfermedad cardiovascular, todas ellas condiciones estrechamente relacionadas con el ictus.

Aunque la relación entre IMC e infarto cerebral puede verse influida por otros factores concomitantes, numerosos estudios han señalado que un IMC elevado incrementa el riesgo de sufrir un evento cerebrovascular, especialmente cuando se combina con otros factores metabólicos adversos.

4.3.5 Tabaquismo y hábitos de vida

El tabaquismo es un factor de riesgo bien establecido para el infarto cerebral. El consumo de tabaco favorece la agregación plaquetaria, el daño endotelial y la aterosclerosis, incrementando el riesgo de formación de trombos. Tanto el tabaquismo activo como la exposición pasiva al humo del tabaco se han relacionado con un aumento significativo del riesgo de ictus.

Otros hábitos de vida, como el sedentarismo y determinados factores socioeconómicos, también influyen en el riesgo de infarto cerebral, ya sea de forma directa o a través de su impacto sobre otros factores clínicos. La consideración conjunta de estos elementos resulta esencial para una evaluación integral del riesgo individual.

5 Objetivos

5.1 Objetivo general

Predecir la probabilidad de ocurrencia de infarto cerebral a partir de variables clínicas y demográficas mediante el uso de modelos estadísticos y de aprendizaje automático aplicados a datos clínicos reales.

5.2 Objetivos específicos

1. Describir las características clínicas y demográficas de la población de estudio, diferenciando entre individuos con y sin infarto cerebral.
2. Evaluar la asociación entre los principales factores de riesgo clínicos y demográficos y la ocurrencia de infarto cerebral mediante contrastes de hipótesis adecuados a la naturaleza de los datos.
3. Analizar las relaciones y correlaciones entre las variables clínicas y metabólicas incluidas en el estudio, con el fin de identificar posibles dependencias entre predictores.
4. Desarrollar un modelo predictivo para estimar la probabilidad de infarto cerebral a partir de las variables disponibles, integrando información clínica relevante.
5. Evaluar y comparar el rendimiento del modelo predictivo mediante métricas de clasificación, valorando su utilidad potencial en un contexto clínico de cribado y prevención.

6 Metodología

6.1 Diseño del estudio

Se realizó un estudio observacional retrospectivo basado en el análisis de un conjunto de datos clínicos secundarios. El diseño del estudio es de tipo analítico y predictivo, con el objetivo de identificar factores asociados a la ocurrencia de infarto cerebral y desarrollar modelos capaces de estimar la probabilidad de aparición del evento a partir de variables clínicas y demográficas.

El análisis combina métodos estadísticos inferenciales clásicos con técnicas de aprendizaje automático supervisado, siguiendo un enfoque reproducible y orientado a la interpretación clínica de los resultados.

6.2 Conjunto de datos

El conjunto de datos analizado contiene información clínica y demográfica de individuos evaluados en relación con la ocurrencia de infarto cerebral. La variable respuesta es la presencia o ausencia de infarto cerebral, codificada como una variable binaria.

Entre las variables explicativas se incluyen la edad, el sexo, el índice de masa corporal, los niveles medios de glucosa en sangre, la presencia de hipertensión arterial y enfermedad cardíaca, el estado civil, el tipo de trabajo, el tipo de residencia y los hábitos de tabaquismo. Estas variables fueron seleccionadas por su relevancia clínica y su asociación documentada con el riesgo de ictus en la literatura científica.

El análisis se llevó a cabo utilizando el lenguaje de programación R (RStudio), garantizando la reproducibilidad mediante el uso de código documentado y paquetes estadísticos ampliamente utilizados en investigación biomédica.

6.3 Preprocesamiento de los datos

Antes del análisis, se realizó un proceso de limpieza y preprocesamiento de los datos. Se evaluó la presencia de valores perdidos en las distintas variables y se excluyeron aquellas observaciones con información incompleta en las variables necesarias para los análisis posteriores.

Las variables categóricas fueron recodificadas y transformadas en factores, asegurando la coherencia de los niveles entre los conjuntos de entrenamiento y prueba. Las variables numéricas fueron convertidas a formato numérico y evaluadas para detectar posibles valores atípicos.

Este preprocesamiento permitió asegurar la calidad de los datos y evitar problemas técnicos durante el entrenamiento y la evaluación de los modelos predictivos.

6.4 Análisis estadístico

Se llevó a cabo un análisis descriptivo inicial de la muestra, resumiendo las variables numéricas mediante medidas de tendencia central y dispersión, y las variables categóricas mediante frecuencias absolutas y relativas.

Para evaluar la asociación entre las variables explicativas y la ocurrencia de infarto cerebral, se aplicaron contrastes de hipótesis. En el caso de variables categóricas, se utilizaron pruebas de independencia basadas en el estadístico chi-cuadrado o la prueba exacta de Fisher cuando las frecuencias esperadas fueron bajas. Para las variables numéricas, se emplearon pruebas no paramétricas para la comparación entre grupos independientes.

Asimismo, se analizaron las correlaciones entre las variables numéricas mediante el coeficiente de correlación de Spearman, con el fin de identificar posibles relaciones y descartar colinealidad elevada entre predictores.

6.5 Modelización predictiva

Para la predicción de la ocurrencia de infarto cerebral se desarrolló un modelo de regresión logística binaria, seleccionado por su amplia utilización en estudios clínicos y por su interpretabilidad.

El conjunto de datos se dividió en un conjunto de entrenamiento y un conjunto de prueba, utilizando una partición aleatoria del 70 % para entrenamiento y del 30 % para evaluación. El modelo fue entrenado utilizando como predictores las variables clínicas y demográficas previamente seleccionadas.

La regresión logística permitió estimar la contribución individual de cada variable al riesgo de infarto cerebral, expresada mediante odds ratios e intervalos de confianza.

6.6 Evaluación del rendimiento del modelo

El rendimiento del modelo predictivo se evaluó utilizando el conjunto de prueba mediante la construcción de matrices de confusión y el cálculo de métricas de clasificación, incluyendo la exactitud, la sensibilidad y la especificidad.

Asimismo, se exploraron distintos umbrales de decisión con el objetivo de analizar el compromiso entre sensibilidad y especificidad, un aspecto especialmente relevante en el contexto de conjuntos de datos desbalanceados. Esta evaluación permitió valorar el comportamiento del modelo en diferentes escenarios clínicos, como el cribado poblacional o la prevención del riesgo individual.

La evaluación del rendimiento mediante métricas como la sensibilidad y la especificidad, junto con la exploración de distintos umbrales de clasificación, está ampliamente recomendada en la literatura metodológica sobre modelización predictiva aplicada, particularmente en problemas de clasificación con clases desbalanceadas (Kuhn & Johnson, 2013). El análisis se realizó íntegramente en RStudio, utilizando funciones y librerías recomendadas en el marco del Máster en Bioestadística y Bioinformática del CEMP, garantizando la coherencia metodológica y la reproducibilidad de los resultados.

7 Resultados

En esta sección se presentan los resultados obtenidos del análisis estadístico y los modelos predictivos desarrollados. A continuación, se detallan las distribuciones de las variables, los contrastes realizados y la interpretación de los modelos.

7.1 Análisis descriptivo

```
##
##   No   Sí
## 4861  249
```

```
## [1] 5110   12
```

7.1.1 Distribución de la edad según infarto cerebral

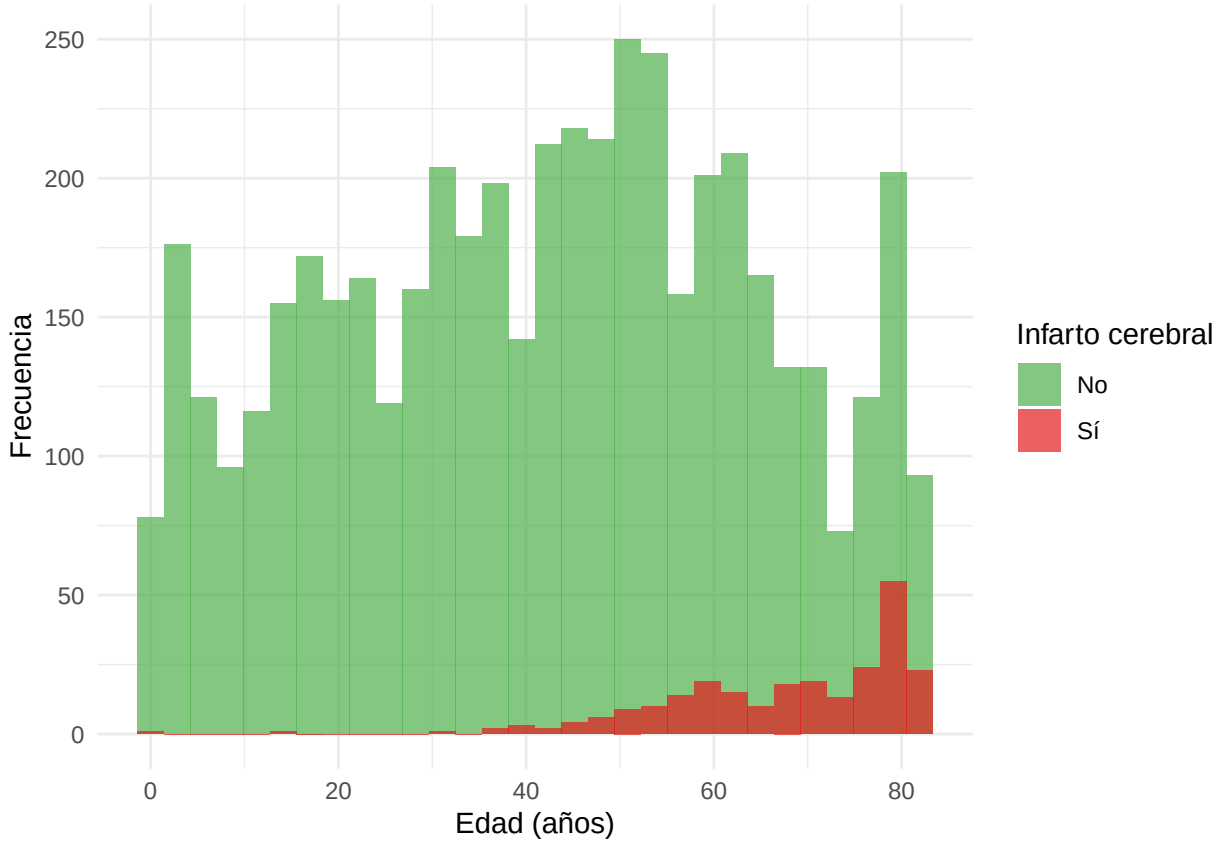


Figure 1: Distribución de la edad según la ocurrencia de infarto cerebral

La Figura 1 muestra la distribución de la edad según la ocurrencia de infarto cerebral. Se observa una mayor concentración de casos de infarto en edades avanzadas, mientras que los individuos sin infarto presentan una distribución más amplia a lo largo del rango de edades.

7.1.2 Comparación de la edad según la presencia de infarto cerebral

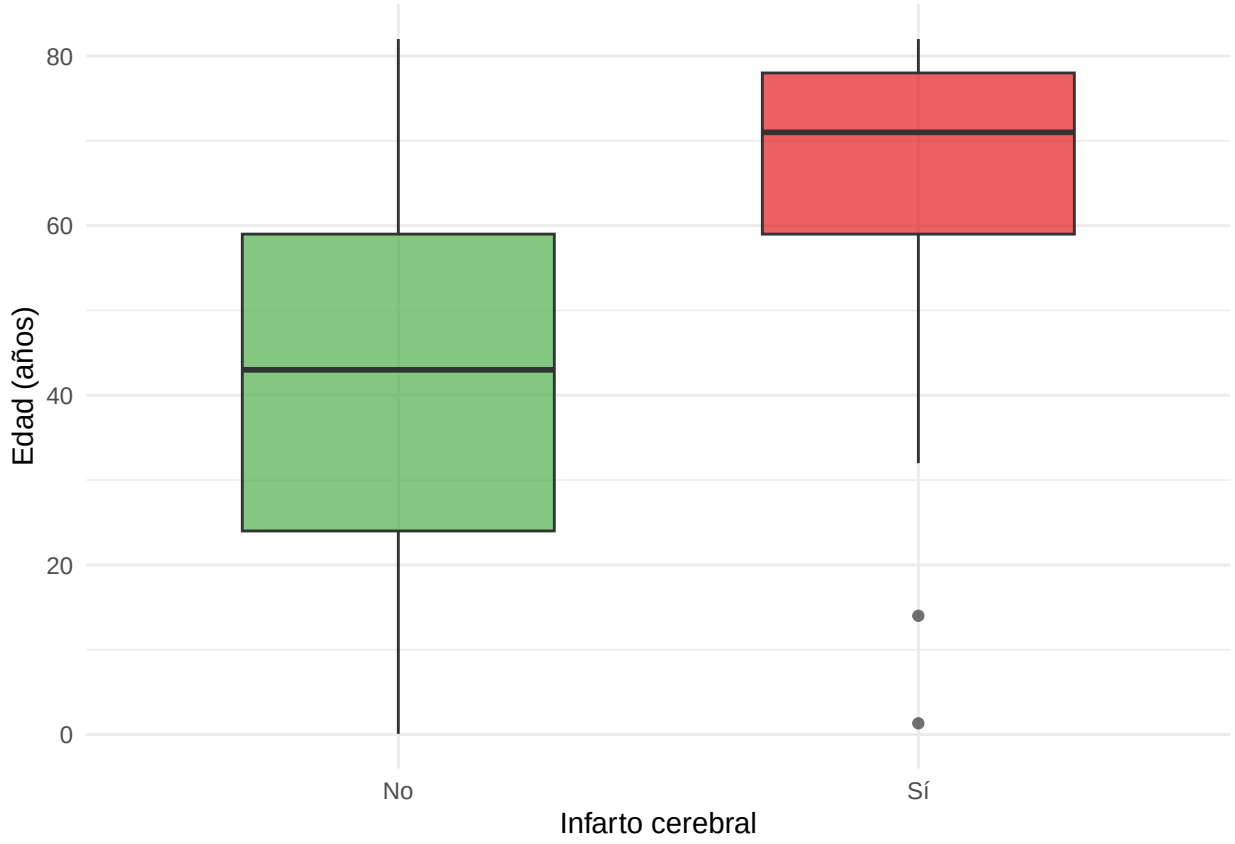


Figure 2: Comparación de la edad según la presencia de infarto cerebral

La Figura 2 compara la distribución de la edad entre individuos con y sin infarto cerebral. Se aprecia que la mediana de edad es claramente superior en el grupo que ha sufrido infarto, lo que refuerza el papel de la edad como factor de riesgo relevante.

7.2 Contrastes de hipótesis

```
##
##   No   Sí
## 4861  249
```

Para la comparación de las variables numéricas entre los grupos de individuos con y sin infarto cerebral se empleó la prueba no paramétrica de Wilcoxon para muestras independientes. La elección de este contraste se fundamenta en varias consideraciones metodológicas. En primer lugar, las variables analizadas (edad, niveles medios de glucosa en sangre e índice de masa corporal) presentan distribuciones que, en estudios clínicos reales, suelen desviarse de la normalidad, especialmente en muestras amplias y heterogéneas como la utilizada en este trabajo. En este contexto, las pruebas paramétricas clásicas, como la t de Student, pueden resultar excesivamente sensibles a la presencia de asimetrías, valores extremos o heterocedasticidad.

Diversos autores recomiendan el uso de pruebas no paramétricas cuando no se puede garantizar el cumplimiento de los supuestos de normalidad o igualdad de varianzas, dado que estas pruebas se basan en el orden de los datos y no en sus valores absolutos, lo que las hace más robustas frente a distribuciones no normales (Conover, 1999; Gibbons & Chakraborti, 2011). En particular, la prueba de Wilcoxon ha sido ampliamente utilizada en estudios biomédicos para la comparación de variables continuas entre dos grupos independientes, mostrando un buen equilibrio entre robustez y potencia estadística.

Además, en el ámbito de la investigación clínica y epidemiológica, es habitual priorizar enfoques conservadores que minimicen el riesgo de conclusiones erróneas derivadas del incumplimiento de supuestos estadísticos, especialmente cuando los resultados van a servir como base para modelos predictivos posteriores. En estudios previos centrados en factores de riesgo de infarto cerebral, variables como la edad, la glucemia o el índice de masa corporal han sido analizadas mediante contrastes no paramétricos similares, reforzando la idoneidad de este enfoque metodológico (Goldstein et al., 2011; O'Donnell et al., 2016).

7.2.1 Comparación de variables categóricas

```
## # A tibble: 7 x 3
##   Variable       Metodo             p_value
##   <chr>          <chr>                <dbl>
## 1 heart_disease  Chi-cuadrado      2.09e-21
## 2 hypertension   Chi-cuadrado      1.66e-19
```

```
## 3 ever_married   Chi-cuadrado      1.64e-14
## 4 smoking_status Chi-cuadrado      2.09e- 6
## 5 work_type      Fisher (simulado) 1.00e- 4
## 6 Residence_type Chi-cuadrado      2.98e- 1
## 7 gender         Fisher (simulado) 5.83e- 1
```

Table 1: Contrastes de asociación entre variables categóricas e infarto cerebral

Variable	Método	p-valor	Significativo
heart_disease	Chi-cuadrado	2.09e-21	Sí
hypertension	Chi-cuadrado	1.66e-19	Sí
ever_married	Chi-cuadrado	1.64e-14	Sí
smoking_status	Chi-cuadrado	2.09e-06	Sí
work_type	Fisher (simulado)	1.00e-04	Sí
Residence_type	Chi-cuadrado	2.98e-01	No
gender	Fisher (simulado)	5.83e-01	No

7.2.2 Comparación de variables numéricas

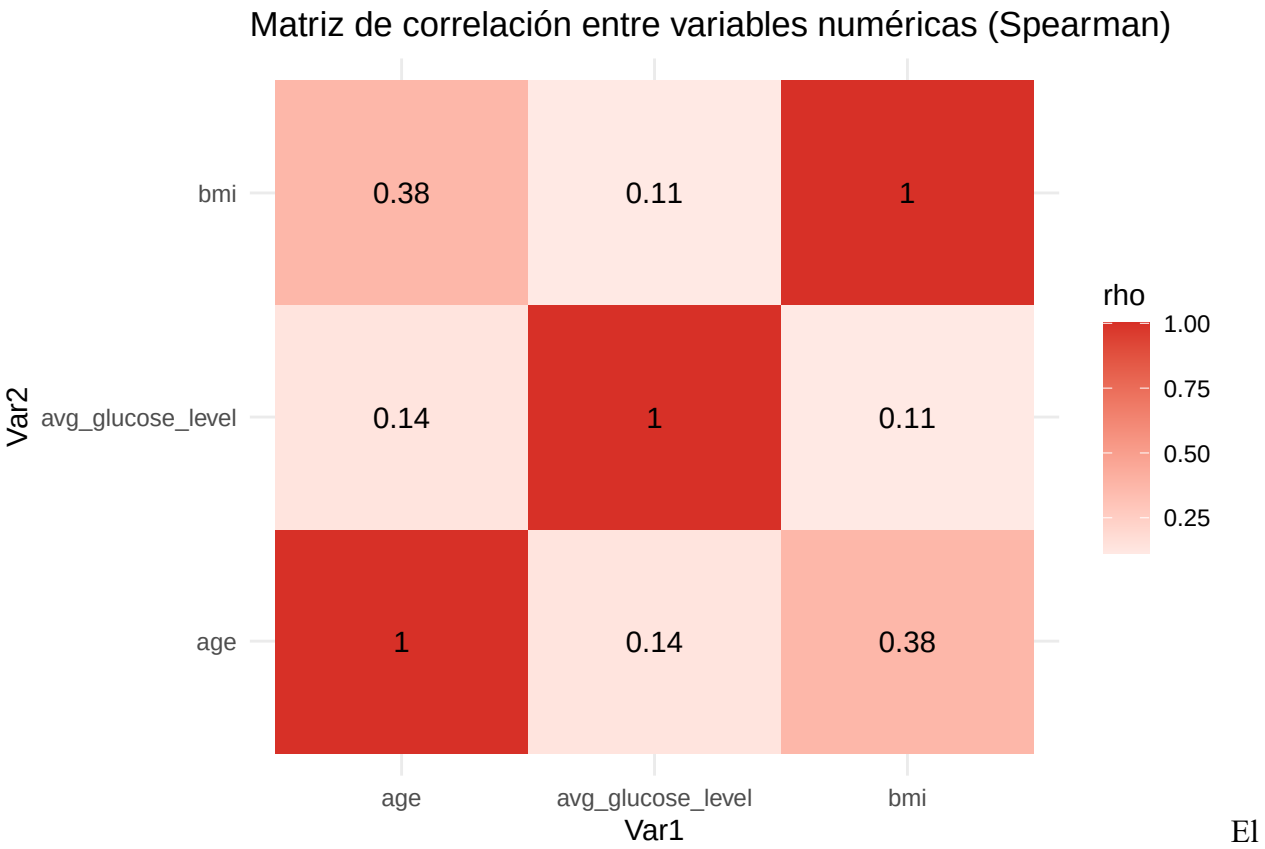
Table 2: Contrastes para variables numéricas entre grupos (No vs Sí)

Variable	Metodo	p-valor	Significativo
Edad	Wilcoxon	3.73e-71	Sí
Glucosa media	Wilcoxon	3.64e-09	Sí
IMC (BMI)	Wilcoxon	1.03e-04	Sí

En las comparaciones entre grupos para variables numéricas se observaron diferencias estadísticamente significativas. La edad mostró la asociación más marcada (Wilcoxon, $p = 3.73 \times 10^{-71}$), con valores superiores en el grupo con infarto cerebral. También se detectaron diferencias para la glucosa media ($p = 3.64 \times 10^{-9}$) y el IMC (BMI) ($p = 1.03 \times 10^{-4}$), lo que sugiere que, en este conjunto de datos, estas variables presentan distribuciones distintas entre individuos con y sin infarto cerebral y podrían aportar información relevante en el desarrollo de modelos predictivos posteriores.

7.3 Análisis de correlación

```
##                          age avg_glucose_level       bmi
## age                1.0000000         0.1408090 0.3756496
## avg_glucose_level 0.1408090         1.0000000 0.1143703
## bmi                0.3756496         0.1143703 1.0000000
```



El análisis de correlación mediante el coeficiente de Spearman mostró asociaciones positivas de magnitud baja a moderada entre las variables numéricas incluidas. La relación más marcada se observó entre edad e IMC ($\rho \approx 0.38$), mientras que las correlaciones entre edad y glucosa media ($\rho \approx 0.14$) y entre IMC y glucosa media ($\rho \approx 0.11$) fueron débiles. En conjunto, no se identificaron correlaciones elevadas, lo que sugiere ausencia de colinealidad severa entre estos predictores numéricos y permite su inclusión conjunta en los modelos predictivos.

7.4 Rendimiento de los modelos predictivos

```
##
##   No   Sí
## 4700  209
```

```
##
## Female   Male  Other
##   2897   2011      1

## [1] 3436   11

## [1] 1473   11

##
## Call:
## glm(formula = stroke ~ age + avg_glucose_level + bmi + gender +
##     hypertension + heart_disease + ever_married + work_type +
##     Residence_type + smoking_status, family = binomial(), data = train)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -7.068e+00  1.091e+00  -6.476 9.39e-11 ***
## age                         7.098e-02  7.592e-03   9.349  < 2e-16 ***
## avg_glucose_level           4.203e-03  1.566e-03   2.684  0.00727 **
## bmi                         1.659e-02  1.394e-02   1.190  0.23395
## genderMale                 -4.242e-02  1.862e-01  -0.228  0.81983
## genderOther                -1.141e+01  2.400e+03  -0.005  0.99620
## hypertension1               4.751e-01  2.125e-01   2.235  0.02539 *
## heart_disease1              3.166e-01  2.510e-01   1.262  0.20711
## ever_marriedYes            -1.401e-01  2.986e-01  -0.469  0.63897
## work_typeGovt_job          -1.189e+00  1.166e+00  -1.019  0.30797
## work_typeNever_worked      -1.128e+01  5.971e+02  -0.019  0.98493
## work_typePrivate           -9.027e-01  1.144e+00  -0.789  0.43017
## work_typeSelf-employed     -1.339e+00  1.169e+00  -1.146  0.25191
## Residence_typeUrban        -1.785e-02  1.805e-01  -0.099  0.92126
## smoking_statusnever smoked -1.535e-01  2.269e-01  -0.677  0.49861
## smoking_statussmokes        3.265e-01  2.707e-01   1.206  0.22783
## smoking_statusUnknown      -3.306e-01  2.915e-01  -1.134  0.25670
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Table 3: (#tab:tab:tabla-logistica)Resultados del modelo de regresión logística (OR e IC 95%)

Variable	OR	IC 95\% (inferior)	IC 95\% (superior)	p-valor
age	1.07	1.06	1.090000e+00	<0.001
avg_glucose_level	1.00	1.00	1.010000e+00	0.007
bmi	1.02	0.99	1.040000e+00	0.234
genderMale	0.96	0.66	1.380000e+00	0.820
genderOther	0.00	NA	1.743289e+207	0.996
hypertension1	1.61	1.05	2.420000e+00	0.025
heart_disease1	1.37	0.83	2.220000e+00	0.207
ever_marriedYes	0.87	0.50	1.620000e+00	0.639
work_typeGovt_job	0.30	0.04	6.300000e+00	0.308
work_typeNever_worked	0.00	0.00	0.000000e+00	0.985
work_typePrivate	0.41	0.06	8.190000e+00	0.430
work_typeSelf-employed	0.26	0.04	5.440000e+00	0.252
Residence_typeUrban	0.98	0.69	1.400000e+00	0.921
smoking_statusnever smoked	0.86	0.55	1.340000e+00	0.499
smoking_statussmokes	1.39	0.81	2.350000e+00	0.228
smoking_statusUnknown	0.72	0.40	1.260000e+00	0.257

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1189.22  on 3435  degrees of freedom
## Residual deviance:  952.28  on 3419  degrees of freedom
## AIC: 986.28
##
## Number of Fisher Scoring iterations: 15
```

Los resultados del modelo de regresión logística muestran que la edad y la hipertensión arterial se asocian de forma significativa con la probabilidad de sufrir un infarto cerebral. En concreto, por cada año adicional de edad, la odds de infarto cerebral aumenta aproximadamente un 7% (OR = 1.07; IC 95\%: 1.06–1.09; p < 0.001). Asimismo, la presencia de hipertensión arterial incrementa significativamente la probabilidad de infarto cerebral (OR = 1.61; IC 95\%: 1.05–2.42; p = 0.025).

Los niveles medios de glucosa en sangre también mostraron una asociación estadísticamente significativa, aunque de menor magnitud (OR = 1.00–1.01; p = 0.007). Por el contrario, variables como el índice de masa corporal, el sexo, el estado civil, el tipo de residencia y los hábitos de tabaquismo no alcanzaron significación estadística en el modelo multivariable.

Algunas categorías, como *work_typeNever_worked*, presentan estimaciones inestables con inter-

valos de confianza amplios, lo que probablemente se deba a un bajo número de observaciones en dichas categorías, una situación habitual en conjuntos de datos clínicos desbalanceados.

```
##     Predicho
## Real   No   Sí
##   No 1407    0
##   Sí   66    0
```

```
##         Metrica Valor
## 1     Exactitud 0.955
## 2  Sensibilidad 0.000
## 3 Especificidad 1.000
```

7.4.1 Evaluación con umbral alternativo

```
##     Predicho
## Real   No   Sí
##   No 1383   24
##   Sí   56   10
```

```
##         Metrica Valor
## 1     Exactitud 0.946
## 2  Sensibilidad 0.152
## 3 Especificidad 0.983
```

Al emplear un umbral de decisión alternativo (0.2), se observa un incremento en la sensibilidad del modelo, permitiendo identificar un mayor número de casos positivos de infarto cerebral. Este aumento se produce a costa de una reducción de la especificidad y de la exactitud global, lo cual es esperable en conjuntos de datos desbalanceados.

Desde un punto de vista clínico, este comportamiento puede resultar aceptable en contextos de cribado, donde se prioriza la detección temprana de posibles casos frente al riesgo de clasificar erróneamente algunos individuos sanos. Por tanto, la elección del umbral debe interpretarse en función del objetivo del modelo y del contexto de aplicación.

8 Discusión

En el presente trabajo se ha llevado a cabo un análisis estadístico y predictivo de la ocurrencia de infarto cerebral a partir de variables clínicas y demográficas, combinando métodos inferenciales clásicos con técnicas de modelización multivariable. Los resultados obtenidos permiten identificar los principales factores asociados al riesgo de infarto cerebral y evaluar el rendimiento de un modelo de regresión logística en un contexto clínico realista.

Un aspecto metodológico relevante a considerar en la interpretación de los resultados es el desbalanceo existente en la variable respuesta del conjunto de datos analizado. La proporción de individuos que han sufrido un infarto cerebral es considerablemente menor que la de aquellos sin evento, una situación habitual en estudios clínicos observacionales y bases de datos poblacionales. Este desbalanceo puede influir tanto en el comportamiento de los modelos predictivos como en la estabilidad de algunas estimaciones.

Asimismo, determinadas variables categóricas incluyen niveles con un número muy reducido de observaciones, como ocurre en el caso de algunas categorías del tipo de trabajo. Esta circunstancia puede dar lugar a estimaciones inestables, errores estándar elevados e intervalos de confianza amplios en los modelos multivariables, sin que ello implique necesariamente una relación clínica relevante. Este fenómeno es bien conocido en regresión logística y debe interpretarse como una limitación inherente a la estructura de los datos más que como un fallo del modelo.

Estas características del conjunto de datos justifican las decisiones metodológicas adoptadas a lo largo del análisis, como el uso de pruebas no paramétricas en los contrastes de hipótesis, el empleo del coeficiente de correlación de Spearman y la exploración de distintos umbrales de clasificación en la evaluación del rendimiento del modelo predictivo. En este sentido, el enfoque seguido prioriza la robustez y la aplicabilidad clínica de los resultados frente a una optimización puramente matemática del modelo.

Uno de los hallazgos más consistentes del estudio es la fuerte asociación entre la edad y la probabilidad de sufrir un infarto cerebral. Tanto en el análisis descriptivo como en los contrastes de hipótesis y el modelo de regresión logística, la edad mostró una relación positiva y estadísticamente significativa con la variable respuesta. En el modelo multivariable, cada incremento anual de edad se asoció con un aumento aproximado del 7 % en la odds de infarto cerebral, resultado que concuerda con lo descrito en estudios epidemiológicos previos, donde la edad se identifica como el principal factor de riesgo no modificable del ictus.

La hipertensión arterial también se confirmó como un factor de riesgo relevante en este estudio. Los individuos con hipertensión presentaron una probabilidad significativamente mayor de infarto cerebral en el modelo de regresión logística, incluso tras ajustar por el resto de variables incluidas.

Este resultado es coherente con la literatura científica, que identifica la hipertensión como el factor de riesgo modificable más importante del ictus, responsable de una proporción sustancial de los casos a nivel poblacional. La inclusión de esta variable en el modelo refuerza su utilidad clínica y su valor predictivo.

Desde el punto de vista clínico, este hallazgo refuerza la importancia del control de la presión arterial como una de las estrategias más efectivas en la prevención primaria del infarto cerebral. Numerosos estudios han demostrado que intervenciones dirigidas a la reducción de la presión arterial se asocian con una disminución significativa del riesgo de ictus, incluso en pacientes sin antecedentes cardiovasculares previos. En este sentido, los resultados obtenidos en el presente trabajo apoyan la relevancia de integrar modelos predictivos basados en factores clínicos fácilmente medibles en programas de prevención y seguimiento en atención primaria y especializada.

En relación con los parámetros metabólicos, los niveles medios de glucosa en sangre mostraron una asociación estadísticamente significativa con la ocurrencia de infarto cerebral, aunque de menor magnitud que la observada para la edad o la hipertensión. Este hallazgo sugiere que alteraciones en el metabolismo de la glucosa, incluso en ausencia de un diagnóstico formal de diabetes, pueden contribuir al riesgo cerebrovascular, tal y como han señalado estudios previos en poblaciones clínicas y comunitarias.

Por el contrario, variables como el índice de masa corporal, el sexo, el estado civil, el tipo de residencia o los hábitos de tabaquismo no alcanzaron significación estadística en el modelo multivariable. Este resultado no implica necesariamente la ausencia de relación con el infarto cerebral, sino que puede deberse a la influencia conjunta de múltiples factores, a posibles efectos de confusión o a la distribución específica de estas variables en el conjunto de datos analizado. En particular, algunas categorías del tipo de trabajo mostraron estimaciones inestables y amplios intervalos de confianza, lo que probablemente se deba al bajo número de observaciones en dichas categorías, una situación habitual en conjuntos de datos clínicos desbalanceados.

En cuanto al rendimiento predictivo del modelo, la evaluación mediante matrices de confusión y métricas de clasificación mostró que el uso del umbral estándar de 0.5 proporciona una elevada especificidad, pero una sensibilidad limitada, lo que implica una menor capacidad para detectar correctamente los casos positivos de infarto cerebral. Dado el carácter desbalanceado del conjunto de datos, este comportamiento era esperable.

La exploración de un umbral alternativo de decisión (0.2) permitió incrementar notablemente la sensibilidad del modelo, identificando un mayor número de individuos con infarto cerebral, a costa de una reducción de la especificidad y de la exactitud global. Desde una perspectiva clínica, este compromiso puede resultar aceptable en contextos de cribado o prevención, donde la prioridad es minimizar los falsos negativos y detectar de forma temprana a individuos con alto riesgo, incluso

si ello conlleva un aumento de falsos positivos.

En conjunto, los resultados obtenidos respaldan la utilidad de los modelos de regresión logística como herramientas predictivas en el ámbito del infarto cerebral, siempre que su interpretación se realice teniendo en cuenta el contexto clínico y las limitaciones inherentes a los datos disponibles. La combinación de análisis estadístico clásico y modelización predictiva aporta una visión integral del problema y permite identificar factores de riesgo relevantes con potencial aplicación en la práctica clínica.

Desde una perspectiva de laboratorio clínico y análisis de datos biomédicos, los resultados obtenidos reflejan situaciones habituales en la práctica real, como el desbalanceo de clases y la presencia de categorías con baja frecuencia, reforzando la aplicabilidad del modelo a escenarios clínicos no ideales.

9 Conclusiones

En este Trabajo Fin de Máster se ha desarrollado un análisis estadístico y predictivo orientado a la identificación de factores asociados a la ocurrencia de infarto cerebral, utilizando un conjunto de datos clínicos con variables demográficas, clínicas y metabólicas. La combinación de análisis descriptivo, contrastes de hipótesis y modelización predictiva ha permitido abordar el problema desde una perspectiva integral y reproducible.

Los resultados obtenidos confirman que la edad constituye el principal factor de riesgo asociado al infarto cerebral, mostrando una relación positiva y altamente significativa en todos los análisis realizados. Asimismo, la hipertensión arterial se identificó como un factor de riesgo relevante en el modelo multivariable, reforzando su papel central en la fisiopatología del ictus y su importancia como diana prioritaria en estrategias de prevención.

Los niveles medios de glucosa en sangre también mostraron una asociación estadísticamente significativa con la probabilidad de infarto cerebral, lo que sugiere que alteraciones metabólicas pueden contribuir al riesgo cerebrovascular incluso en ausencia de un diagnóstico explícito de diabetes. Por el contrario, otras variables como el índice de masa corporal, el sexo, el tipo de residencia o los hábitos de tabaquismo no alcanzaron significación estadística en el modelo ajustado, lo que pone de manifiesto la complejidad multifactorial del infarto cerebral y la necesidad de interpretar estos resultados en conjunto.

Desde el punto de vista predictivo, el modelo de regresión logística mostró un comportamiento coherente con la naturaleza desbalanceada del conjunto de datos. El uso del umbral de decisión estándar priorizó la especificidad, mientras que la exploración de un umbral alternativo permitió mejorar la sensibilidad del modelo, incrementando la detección de casos positivos. Este resultado subraya la importancia de adaptar el umbral de clasificación al contexto clínico concreto, especialmente en escenarios de cribado o prevención, donde la reducción de falsos negativos puede ser prioritaria.

Entre las principales limitaciones del estudio se encuentran el desbalanceo de la variable respuesta, la presencia de categorías con escaso número de observaciones y la ausencia de información clínica adicional que podría mejorar el rendimiento predictivo del modelo. No obstante, estas limitaciones reflejan situaciones habituales en datos clínicos reales y refuerzan el valor práctico del enfoque adoptado.

En conclusión, este trabajo demuestra el potencial de los métodos estadísticos y de aprendizaje automático, en particular la regresión logística, para apoyar la identificación de factores de riesgo y la predicción del infarto cerebral. Los resultados obtenidos pueden servir como base para futuros estudios que incorporen nuevos predictores, técnicas de modelización más complejas o validaciones

externas, contribuyendo así al avance de la medicina predictiva y preventiva en el ámbito de la salud cerebrovascular.

En este contexto, el presente Trabajo Fin de Máster pone de manifiesto el papel fundamental de la bioestadística y el análisis de datos clínicos como herramientas clave para la medicina basada en la evidencia. La correcta aplicación e interpretación de métodos estadísticos y modelos predictivos permite no solo describir asociaciones, sino también apoyar la toma de decisiones clínicas y el diseño de estrategias preventivas más eficientes, contribuyendo a una atención sanitaria más personalizada y orientada al riesgo.

10 Bibliografía

Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). John Wiley & Sons.

Feigin, V. L., Norrving, B., & Mensah, G. A. (2017). Global burden of stroke. *Circulation Research, 120*(3), 439–448. https://doi.org/10.1161/CIRCRESAHA.116.308413

Gibbons, J. D., & Chakraborti, S. (2011). *Nonparametric statistical inference* (5th ed.). Chapman & Hall/CRC.

Goldstein, L. B., Bushnell, C. D., Adams, R. J., Appel, L. J., Braun, L. T., Chaturvedi, S., Creager, M. A., Culebras, A., Eckel, R. H., Hart, R. G., Hinchey, J. A., Howard, V. J., Jauch, E. C., Levine, S. R., Meschia, J. F., Moore, W. S., Nixon, J. V., & Pearson, T. A. (2011). Guidelines for the primary prevention of stroke. *Stroke, 42*(2), 517–584. https://doi.org/10.1161/STR.0b013e3181fcb238

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

Kannel, W. B., Wolf, P. A., Verter, J., & McNamara, P. M. (1970). Epidemiologic assessment of the role of blood pressure in stroke. *Journal of the American Medical Association, 214*(2), 301–310. https://doi.org/10.1001/jama.1970.03180020051010

O'Donnell, M. J., Chin, S. L., Rangarajan, S., Xavier, D., Liu, L., Zhang, H., Rao-Melacini, P., Zhang, X., Pais, P., Agapay, S., Lopez-Jaramillo, P., Damasceno, A., Langhorne, P., McQueen, M. J., Rosengren, A., Dehghan, M., Hankey, G. J., Dans, A. L., Elsayed, A., … Yusuf, S. (2016). Global and regional effects of potentially modifiable risk factors associated with acute stroke. *The Lancet, 388*(10046), 761–775. https://doi.org/10.1016/S0140-6736(16)30506-2

Powers, W. J., Rabinstein, A. A., Ackerson, T., Adeoye, O. M., Bambakidis, N. C., Becker, K., Biller, J., Brown, M., Demaerschalk, B. M., Hoh, B., Jauch, E. C., Kidwell, C. S., Leslie-Mazwi, T. M., Ovbiagele, B., Scott, P. A., Sheth, K. N., Southerland, A. M., Summers, D. V., & Tirschwell, D. L. (2018). Guidelines for the early management of patients with acute ischemic stroke. *Stroke, 49*(3), e46–e110. https://doi.org/10.1161/STR.0000000000000158

Rothwell, P. M. (2007). Prevention of stroke in patients with transient ischemic attack and minor stroke. *The Lancet, 370*(9589), 1432–1442. https://doi.org/10.1016/S0140-6736(07)61448-2

Steyerberg, E. W. (2019). *Clinical prediction models* (2nd ed.). Springer.

World Health Organization. (2023). *Stroke*. https://www.who.int/news-room/fact-sheets/detail/stroke

Heo, J., Yoon, J. G., Park, H., Kim, Y. D., Nam, H. S., & Heo, J. H. (2019). Machine learning–based model for prediction of outcomes in acute stroke. Stroke, 50(5), 1263–1265. https://doi.org/

10.1161/STROKEAHA.118.024293

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.

11 Anexo: Código fuente

El presente Trabajo Fin de Máster se ha desarrollado siguiendo un enfoque reproducible, con el objetivo de garantizar la transparencia metodológica y facilitar la replicación de los resultados obtenidos.

Todos los análisis estadísticos, las visualizaciones, los contrastes de hipótesis, los análisis de correlación y los modelos predictivos presentados en esta memoria han sido implementados en el lenguaje de programación **R**, utilizando el entorno **RStudio** y paquetes ampliamente empleados en el ámbito de la bioestadística y el análisis de datos clínicos.

Con el fin de asegurar el acceso al código fuente y permitir la reproducibilidad del estudio, el código completo utilizado en este trabajo se encuentra disponible en un repositorio público de GitHub, accesible a través del siguiente enlace:

https://github.com/virvisco/TFM_infarto_cerebral

En dicho repositorio se incluye el archivo principal en formato `.Rmd`, junto con los scripts auxiliares necesarios para la ejecución del análisis y la generación de los resultados presentados en la memoria. La estructura del repositorio permite seguir de forma clara el flujo de trabajo realizado, desde el preprocesamiento de los datos hasta la evaluación del rendimiento de los modelos predictivos.

El uso de GitHub como plataforma de alojamiento del código facilita el control de versiones, la trazabilidad de los cambios realizados durante el desarrollo del proyecto y la posible reutilización del código en futuros trabajos o estudios relacionados. Este anexo garantiza el cumplimiento de los criterios de reproducibilidad exigidos en el Máster en Bioestadística y Bioinformática del Centro Europeo de Másteres y Posgrados y refuerza el carácter técnico y aplicado del trabajo realizado.