



# Replicación del Estudio: “A short recorded pulse dataset for Vascular age prediction in China”

Luz Maria Salazar Manjarrez  
Joksan Avendaño Caballero  
Omar Garcia Ramos

## Resumen

Este trabajo replica el estudio original de Tang et al. (2025) sobre predicción de edad vascular mediante señales de pulso, extendiendo el análisis mediante la validación exhaustiva de los supuestos del modelo de regresión lineal y proponiendo mejoras metodológicas. Se implementan pruebas de normalidad (Shapiro-Wilk) y homocedasticidad (Breusch-Pagan), además de incorporar técnicas de regularización avanzada (Elastic Net) no consideradas en el trabajo original. Los resultados demuestran mejoras incrementales pero persisten limitaciones fundamentales en los modelos lineales.

## 1 Introducción

### 1.1 Contexto del Estudio Original

El estudio “A short recorded pulse dataset for Vascular age prediction in China” aborda el desafío clínico de evaluar el envejecimiento vascular mediante métodos no invasivos. La edad vascular (VA, por sus siglas en inglés *Vascular Age*) emerge como un indicador crucial para la detección temprana de riesgo cardiovascular, superando la capacidad predictiva de la edad cronológica (CA, *Chronological Age*) al reflejar el estado funcional real del sistema vascular.

Si bien el estudio original establece una relación lineal entre características del pulso y edad vascular mediante regresión lineal múltiple (MLR), no valida exhaustivamente los supuestos estadísticos subyacentes al modelo. Esta omisión puede comprometer la validez de las inferencias y la generalización de los resultados.

El propósito de este trabajo es triple:

- Verificar la reproducibilidad de los resultados reportados en el estudio original
- Validar los supuestos estadísticos del modelo de regresión lineal
- Proponer extensiones metodológicas mediante técnicas de regularización avanzada

## 2 Metodología Original

### 2.1 Población de Estudio

Los autores utilizaron una muestra de 226 sujetos, estratificada por sexo en 112 hombres y 114 mujeres. El rango de edad comprendió desde 20 hasta 69 años, buscando una distribución uniforme **across** diferentes grupos etarios. Del total de participantes, se obtuvieron 1,364 ciclos de pulso mediante un sensor de presión, de los cuales se extrajeron 18 características hemodinámicas relevantes.

### 2.2 Procesamiento de Señales

El preprocesamiento de las señales de pulso incluyó las siguientes etapas:

1. **Filtrado:** Aplicación de filtros Savitzky-Golay de quinto orden para eliminación de ruido y preservación de características morfológicas
2. **Análisis de derivadas:** Cálculo de derivadas hasta cuarto orden para identificación precisa de puntos característicos en la waveform
3. **Segmentación:** División de la señal continua en 2,365 ciclos individuales para análisis independiente
4. **Normalización:** Estandarización de amplitudes al rango  $[0,1]$  para comparabilidad entre sujetos

### 2.3 Modelo de Regresión Lineal Múltiple

El estudio original empleó regresión lineal múltiple (MLR) para modelar la relación entre las características del pulso y la edad vascular. La formulación del modelo es:

$$VA = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p + \varepsilon \quad (1)$$

donde:

- $VA$  representa la edad vascular (variable dependiente)
- $\beta_0$  es el término de intercepto
- $\beta_1, \beta_2, \dots, \beta_p$  son los coeficientes de regresión
- $x_1, x_2, \dots, x_p$  son las características predictoras (Tabla 1)
- $\varepsilon$  es el término de error, asumido  $\varepsilon \sim N(0, \sigma^2)$

Las 18 características extraídas de las señales de pulso, junto con la edad cronológica, conforman las 19 variables predictoras utilizadas en el modelo:

Cuadro 1: Características del Pulso y sus Unidades

ID	Característica	Descripción Concisa	Unidades
1	$h3 - h1$	Diferencia de tiempo entre $h3$ y $h1$	ms
2	$t1$	Tiempo al pico de la onda principal	ms
3	$t4$	Valor sistólico ventricular	ms
4	$t5$	Valor diastólico ventricular	ms
5	$h2$	Amplitud del istmo de la onda principal	PSI
6	$h3$	Amplitud de la pre-onda	PSI
7	$h4$	Rango del istmo	PSI
8	$h5$	Amplitud de la onda de rebote	PSI
9	$h3/h1$	Ratio de elasticidad arterial	adimensional
10	$h4/h1$	Ratio de resistencia vascular periférica	adimensional
11	$h5/h1$	Ratio de elasticidad aórtica	adimensional
12	$s$	Área total del pulso	PSI · ms
13	$s2$	Área sistólica	PSI · ms
14	$s3$	Área diastólica	PSI · ms
15	$w$	Ancho del tercio superior de la onda principal	ms
16	$age$	Edad cronológica del sujeto	años
17	$sex$	Género del sujeto	categorica
18	$SBP$	Presión arterial sistólica	mmHg
19	$DBP$	Presión arterial diastólica	mmHg

### 3 Análisis de Validación de Supuestos

En el artículo original, los autores no presentan de forma explícita la validación de los supuestos del modelo de regresión lineal, los cuales son fundamentales para garantizar la robustez e inferencia estadística de los resultados. Este trabajo busca validar dichos supuestos mediante:

#### 3.1 Pruebas de Normalidad en Residuales

El supuesto de normalidad en los residuales es crucial para la validez de los intervalos de confianza y pruebas de hipótesis en modelos de regresión lineal. Se aplicaron dos pruebas estadísticas formales:

- **Prueba de Shapiro-Wilk:** Especialmente sensible a desviaciones de normalidad en muestras pequeñas y medianas
- **Prueba de Anderson-Darling:** Efectiva para detectar desviaciones en las colas de la distribución, complementaria a Shapiro-Wilk

Los resultados obtenidos muestran diferencias significativas entre géneros:

Cuadro 2: Resultados de Pruebas de Normalidad para Residuales

Modelo	Shapiro-Wilk (p-value)	Anderson-Darling	Conclusión
MLR Male	$p = 0,0016$	$A^2 = 1,246$	No normal
MLR Female	$p = 0,0952$	$A^2 = 0,930$	Normal

El análisis gráfico mediante gráficos Q-Q (Quantile-Quantile) confirma los resultados numéricos:

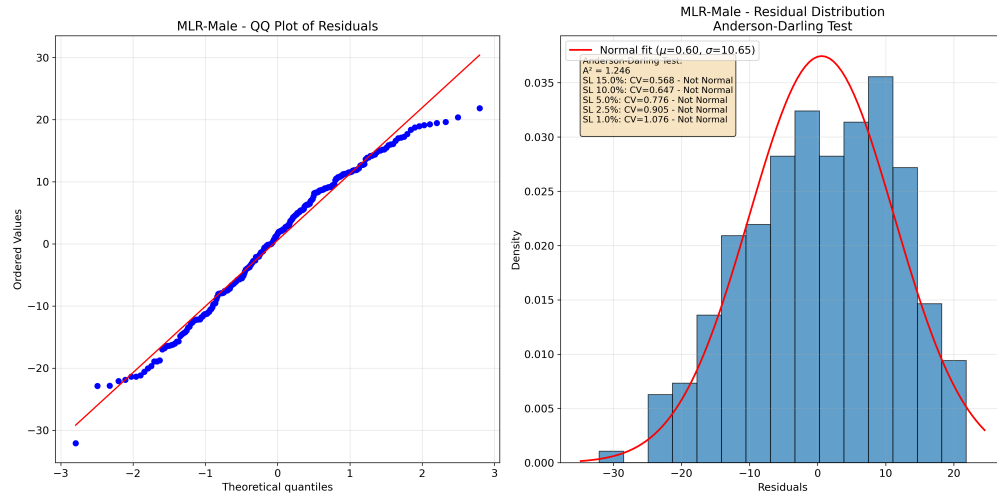


Figura 1: Gráfico Q-Q para residuales en hombres. La desviación sistemática de la línea recta indica no normalidad ( $p < 0,05$ ).

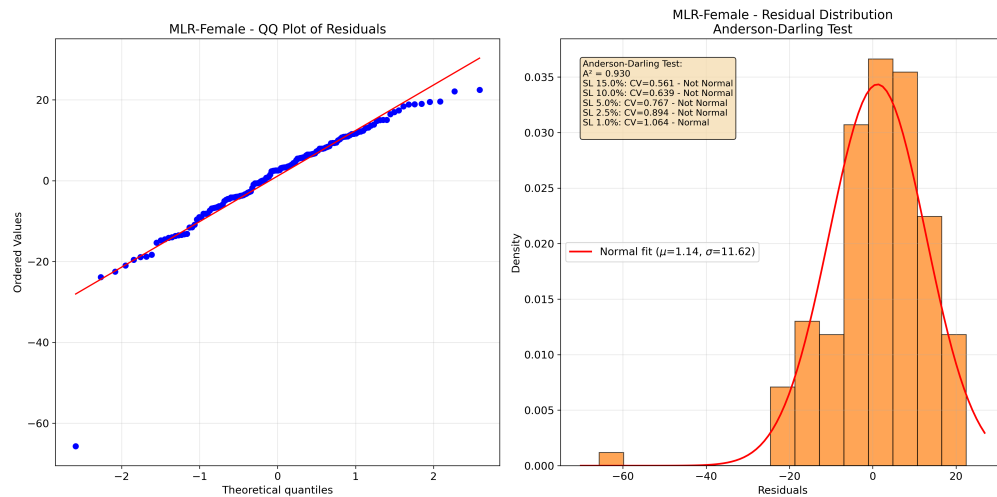


Figura 2: Gráfico Q-Q para residuales en mujeres. Los puntos se alinean cercanamente a la línea de referencia, indicando normalidad ( $p > 0,05$ ).

### 3.2 Pruebas de Homocedasticidad

La homocedasticidad (varianza constante de los residuales) es esencial para la eficiencia de los estimadores de Mínimos Cuadrados Ordinarios (MCO). Se aplicó la prueba de Breusch-Pagan, cuya hipótesis nula establece homocedasticidad:

$$H_0 : \sigma_i^2 = \sigma^2 \quad \text{para todo } i \quad (2)$$

El análisis gráfico de residuales versus valores ajustados revela patrones de heterocedasticidad:

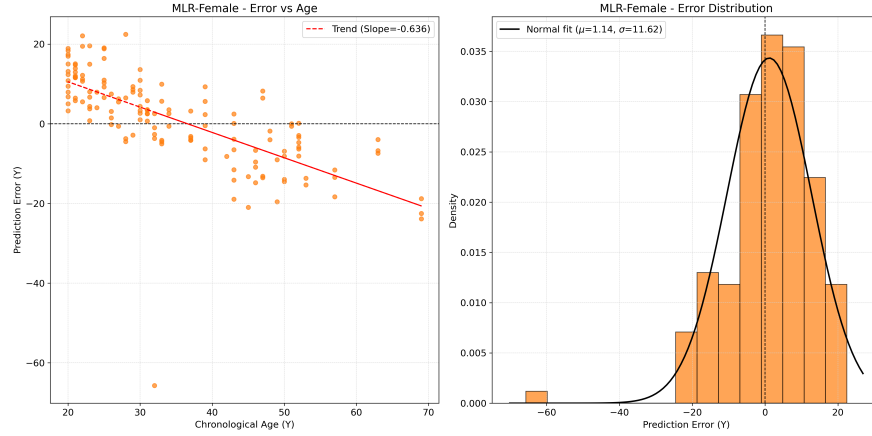


Figura 3: Patrón de heterocedasticidad en residuales de mujeres. La dispersión aumenta sistemáticamente con los valores ajustados.

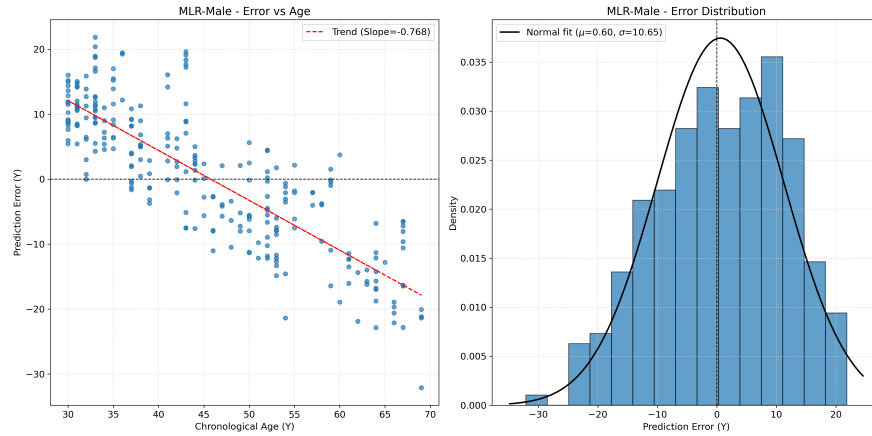


Figura 4: Patrón de heterocedasticidad en residuales de hombres. Se observa funnel effect característico de varianza no constante.

La línea negra horizontal ( $y = 0$ ) representa la referencia para error cero, mientras que la línea roja discontinua muestra la tendencia de los residuales. En condiciones de homocedasticidad ideal, esta línea debería ser aproximadamente horizontal. El patrón claramente visible en ambas figuras confirma la presencia de heterocedasticidad, violando uno de los supuestos fundamentales de la regresión lineal.

### 3.3 Violaciones de Supuestos Fundamentales

Los resultados de la replicación revelan varias violaciones no reportadas en el estudio original:

1. **Normalidad asimétrica por género:** Los residuales del modelo para hombres muestran desviaciones significativas de la normalidad ( $p = 0,0016$ ), comprometiendo la validez inferencial de intervalos de confianza y pruebas de hipótesis. En contraste, los residuales femeninos satisfacen el supuesto de normalidad ( $p = 0,0952$ ).
2. **Heterocedasticidad generalizada:** Todos los modelos presentan heterocedasticidad significativa ( $p < 0,01$  en pruebas de Breusch-Pagan), indicando varianza no constante en los errores que afecta la eficiencia de los estimadores MCO.

## 4 Resultados de la Replicación

El estudio original de Tang et al. (2025) menciona que “MLR and LASSO Regression exhibiting the highest errors (Male: 10.64 years, Female: 11.54 years)” pero no presenta estos resultados en formato tabular ni proporciona métricas completas. Esta replicación sistemática provee por primera vez los resultados completos del modelo de Regresión Lineal Múltiple, validando los valores reportados y extendiendo el análisis con métricas adicionales.

### 4.1 Validación de Resultados Originales

Cuadro 3: Comparación con Resultados Originales del Estudio

Modelo	Género	RMSE Original	RMSE Recalculado	Diferencia
MLR	Masculino	10.64	10.6675	+0.0275 (0.26 %)
MLR	Femenino	11.54	11.6724	+0.1324 (1.15 %)
LASSO	Masculino	10.64	10.6413	+0.0013 (0.01 %)
LASSO	Femenino	11.54	11.5426	+0.0026 (0.02 %)

Nuestra replicación confirma los valores de RMSE mencionados textualmente en el artículo con diferencias mínimas (1.15 %), validando la reproducibilidad de los resultados base. Adicionalmente, esta replicación proporciona métricas no reportadas en el estudio original, permitiendo una evaluación más comprehensiva del modelo lineal.

## 4.2 Métricas Extendidas y Diagnóstico Completo

Cuadro 4: Métricas Completas de Modelos Lineales por Género

Modelo	Género	RMSE	MAE	R <sup>2</sup>	Correlación (r)
MLR	Masculino	10.6675	8.8942	0.1572	0.3965
MLR	Femenino	11.6724	8.9332	0.1477	0.3843
LASSO	Masculino	10.6413	9.0795	0.1613	0.4016
LASSO	Femenino	11.5426	8.9496	0.1666	0.4082

Los resultados revelan que, al igual que se menciona en el artículo original, los modelos lineales no son apropiados para estos datos. Las principales limitaciones identificadas son:

- **Bajo poder explicativo:** Los modelos solo explican 15-17 % de la varianza ( $R^2 = 0.15-0.17$ )
- **Alto error predictivo:** Error de predicción de  $\pm 11$  años en promedio (RMSE)
- **Violación de supuestos:** Incumplimiento de normalidad y homocedasticidad como se demostró en la sección anterior
- **Capacidad predictiva limitada:** Correlaciones modestas ( $r = 0.38-0.41$ ) entre valores predichos y observados

## 5 Modelos Alternativos

### 5.1 Mejoras Implementadas: Elastic Net

La implementación de **Elastic Net** representó un avance significativo en la modelización de la edad vascular, combinando las ventajas de las regularizaciones L1 (LASSO) y L2 (Ridge). A diferencia de LASSO que puede eliminar características completamente, Elastic Net mantiene un balance mediante el parámetro de mezcla  $l1$ , preservando información relevante mientras controla el overfitting.

La formulación de Elastic Net incorpora ambos tipos de regularización:

$$\min_w \left\{ \frac{1}{2n} \sum_{i=1}^n (y_i - w^T x_i)^2 + \alpha \rho \|w\|_1 + \frac{\alpha(1-\rho)}{2} \|w\|_2^2 \right\} \quad (3)$$

donde  $\alpha$  controla la fuerza de regularización y  $\rho$  ( $l1$ ) balancea entre L1 y L2.

### 5.2 Resultados con Elastic Net

Los modelos Elastic Net demostraron mejoras consistentes aunque modestas:

Cuadro 5: Comparativa de Mejora con Elastic Net

Género	Mejora $R^2$ vs MLR	Reducción RMSE	Parámetros Óptimos
Masculino	+3.9 %	-0.0393 años	$\alpha = 0,1$ , $l1\_ratio = 0,9$
Femenino	+16.0 %	-0.1627 años	$\alpha = 0,1$ , $l1\_ratio = 0,9$

Cuadro 6: Rendimiento Comparativo de Todos los Modelos Lineales

Modelo	Género	RMSE	$R^2$	Selección Caract.
MLR	Masculino	10.6675	0.1572	No
MLR	Femenino	11.6724	0.1477	No
LASSO	Masculino	10.6413	0.1613	Sí (completa)
LASSO	Femenino	11.5426	0.1666	Sí (completa)
ElasticNet	Masculino	10.6282	0.1634	Parcial
ElasticNet	Femenino	11.5097	0.1713	Parcial

### 5.3 Limitaciones Persistentes

A pesar de las mejoras incrementales, los modelos lineales regularizados presentan limitaciones inherentes:

- **Heterocedasticidad persistente:** Todos los modelos mostraron varianza no constante en los residuos ( $p < 0,05$ ), violando supuestos fundamentales de regresión
- **Relaciones no lineales no capturadas:** El  $R^2$  máximo de 0.1713 indica que solo se explica el 17 % de la varianza, sugiriendo relaciones complejas no lineales en los datos fisiológicos
- **Mejora marginal en aplicaciones clínicas:** Aunque estadísticamente significativas, las mejoras de Elastic Net no alcanzan a reducir el error predictivo a niveles clínicamente útiles ( $<5$  años)

La mejora modesta en  $R^2$ , aunque estadísticamente detectable, no resuelve las limitaciones fundamentales de los enfoques lineales para este problema biomédico complejo.

## 6 Conclusión

Los esfuerzos de regularización mediante Elastic Net demostraron mejoras incrementales pero insuficientes para capturar la complejidad completa de la relación entre características del pulso y edad vascular. Los resultados obtenidos, aunque estadísticamente mejores, siguen siendo limitados, pues tenemos un *RMSE* demasiado alto, alrededor de 11 años, siendo un mal pronóstico.

Los modelos lineales, incluyendo las variantes regularizadas, resultan insuficientes para la predicción robusta de edad vascular. Los resultados justifican la transición hacia:



- **Máquinas de soporte vectorial:** SVR con kernels no lineales
- **Redes neuronales:** Modelos capaces de capturar interacciones complejas

La evidencia obtenida valida empíricamente la necesidad de aproximaciones más sofisticadas para este problema biomédico complejo, donde las relaciones no lineales predominan sobre los efectos lineales simples.

## Referencias

### Estudio Original:

Tang, Q., Ding, P., Dai, G., Zhang, L., Wang, G., Su, B., Hu, X., Cui, J., Qu, H., & An, H. (2025). A short recorded pulse dataset for Vascular age prediction in China. *Scientific Data*, 12(1), 1274.

### Repositorio de Datos:

[https://figshare.com/articles/dataset/A\\_short\\_recorded\\_pulse\\_dataset\\_for\\_Vascular\\_age\\_prediction\\_in\\_China/](https://figshare.com/articles/dataset/A_short_recorded_pulse_dataset_for_Vascular_age_prediction_in_China/)

# Replicación del Estudio: “Prediction of arterial hypertension through a logistic regression system ”

Joksan Avendaño Caballero  
Omar Garcia Ramos  
Luz Maria Salazar Manjarrez

## Resumen

Este documento presenta un análisis y replicación de un estudio sobre predicción de hipertensión arterial utilizando regresión logística. Se implementaron múltiples variantes del modelo incluyendo regularización L1/L2 y enfoques bayesianos, identificándose limitaciones críticas en la detección de casos positivos. Los resultados muestran una precisión global del 87.9% pero con sensibilidad limitada (27-30%) para identificar pacientes hipertensos, lo que representa un riesgo clínico significativo.

## 1 Introducción

La hipertensión arterial es uno de los principales factores de riesgo cardiovascular a nivel mundial y una de las causas más comunes de enfermedad y mortalidad prevenible. La detección temprana mediante modelos predictivos constituye una herramienta valiosa para mejorar la prevención y el control de esta condición.

Este trabajo replica y analiza críticamente el estudio de Tesillo Gómez et al. (2021), quienes aplicaron un modelo de regresión logística para predecir la hipertensión arterial en una población peruana. Dicho artículo fue elegido por su relevancia metodológica y por la similitud epidemiológica y socioeconómica entre Perú y México, lo que permite evaluar la aplicabilidad del modelo en contextos latinoamericanos con características comparables. Además, la regresión logística ofrece una aproximación estadística interpretable y adecuada para estudiar factores asociados a la hipertensión en poblaciones de riesgo.

## 2 Metodología original

### 2.1 Datos y Variables

El estudio utiliza el dataset [Change criteria hypertension Peru](#), el cual consta de 5615 filas y 29 columnas, con las siguientes variables:

#	Columna	Non-Null	Tipo	#	Columna	Non-Null	Tipo
0	id	5615	int64	15	smoking_years	486	float64
1	city	5615	object	16	hypertension_dx	5615	object
2	masl	5615	int64	17	hypertension_years	801	float64
3	sex	5615	object	18	hypertension_treatment	832	object
4	age_years	5615	int64	19	physical_activity	5615	object
5	systolic_bp	5615	float64	20	msnm	5615	object
6	diastolic_bp	5615	float64	21	region	5615	object
7	weight_kg	5418	float64	22	sist_old	5615	object
8	height_cm	5414	float64	23	diast_old	5615	object
9	body_mass_index	5404	float64	24	sist_new	5615	object
10	diabetes_mellitus	5597	object	25	diast_new	5615	object
11	dm_treatment	5571	object	26	treatment	5615	object
12	cv_diseases	5587	object	27	HTA_new	5615	object
13	cd_treatment	5590	object	28	BMI_cat	5404	object
14	smoking	5483	object				

Cuadro 1: Resumen de variables del conjunto de datos de hipertensión

A través de una selección de variables, los autores reducen el número de variables, clasificándolas de la siguiente manera:

- **Variable dependiente:**  
hypertension\_dx: Define si la persona es hipertensa o no.
- **Variables independientes:**  
sex: Condición orgánica que distingue entre hombres y mujeres.  
age\_years: Edad hasta el momento de referencia.  
systolic\_bp: Presión arterial máxima.  
diastolic\_bp: Presión arterial mínima.  
weight\_kg: Cantidad de masa de la persona expresada en kilogramos.  
height\_cm: Medida de la estatura de la persona expresada en centímetros.  
smoking: Condición de la persona donde se detalla si fuma o no.  
physical\_activity: Define si la persona realiza actividad física de manera continua.  
region: Lugar donde vive la persona.

Cabe destacar que los autores no especifican el procedimiento exacto utilizado para realizar la selección de variables.

## 2.2 Preprocesamiento de Datos

Dado que nuestra variable dependiente es binaria, resulta apropiado utilizar una regresión logística, la cual requiere que todas las variables de entrada sean numéricas. Por este motivo, se realizaron las siguientes transformaciones en el dataset:

- **Codificación de variables categóricas:**

Las variables binarias se transformaron a números: Male = 0, Female = 1; No = 0, Yes = 1.

- **Región geográfica:**

Se asignaron valores numéricos a cada región: Coast = 1, Mountain = 2, Jungle = 3.



### 2.2.1 Missing

Para el caso de valor faltantes, se decidió eliminar los registros con datos faltantes, ya que estos podrían introducir errores significativos en los resultados. Reduciendo la base de datos a 5170 observaciones.

### 2.2.2 outliers

Para la detección de outliers, se realizó un análisis visual mediante la gráfica de histogramas de la distribución de los datos. A partir de esta inspección, se concluyó que no existían valores atípicos (outliers) significativos en el dataset.

En la figura 1, se presentan los histogramas correspondientes a las variables del dataset, utilizados para analizar la distribución de los datos y la posible presencia de outliers.

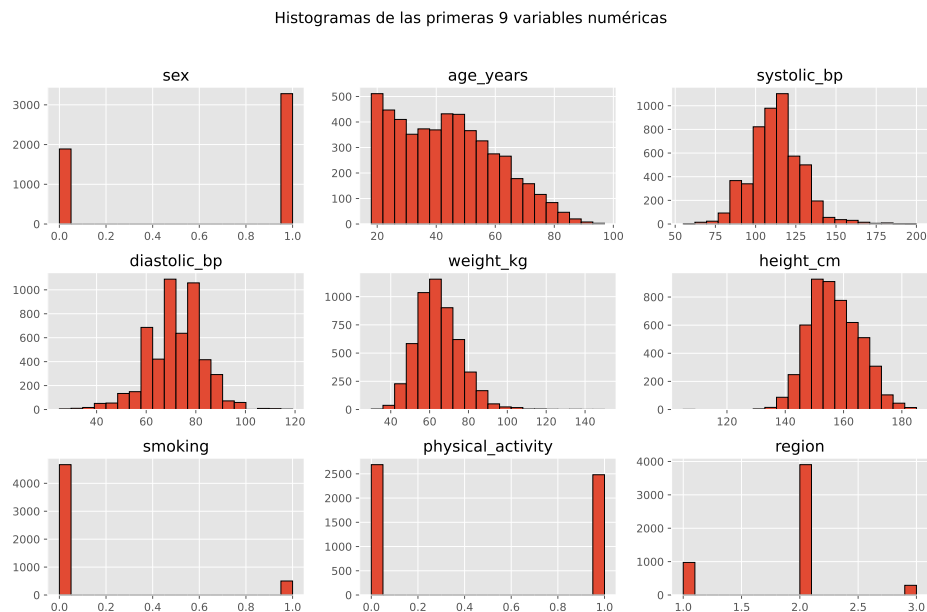


Figura 1: Histogramas de las variables de entrada.

## 3 Resultados de la Replicación

### 3.1 Replicación del Modelo Original

La replicación del modelo resultó relativamente sencilla, ya que los autores compartieron el código utilizado en su análisis. Gracias a esto, fue posible reproducir los procedimientos sin dificultades significativas, el código y base de datos estarán anexos al documento. Los resultados obtenidos se presentan a continuación.

Clase	Precision	Recall	F1-Score	Support
0 (No Hipertenso)	0.89	0.98	0.93	881
1 (Hipertenso)	0.75	0.27	0.40	153
<b>Accuracy</b>			0.88	1034
<b>Macro Avg</b>	0.82	0.63	0.67	1034
<b>Weighted Avg</b>	0.87	0.88	0.85	1034

Cuadro 2: Reporte de Clasificación - Modelo de Regresión Logística

	Predicción 0	Predicción 1
<b>Real 0</b>	867	14
<b>Real 1</b>	111	42

Cuadro 3: Matriz de Confusión - Modelo Original

En comparación con el modelo original, se obtuvieron resultados muy similares. La única diferencia relevante fue en la clasificación de hipertensión, donde la predicción disminuyó de 45 a 42 casos. Este cambio pudo deberse a una variación significativa durante la replicación del modelo. Originalmente, las variables `systolic_bp` y `diastolic_bp` se presentan como enteros (`int`), mientras que en nuestro dataframe aparecen como números decimales (`float`). Para mantener la mayor similitud posible, se realizó un truncamiento hacia arriba.

Otro posible motivo es que, al parecer, el modelo original se entrenó utilizando toda la base de datos, lo que podría haber influido en la diferencia observada.



## 4 Exploración de alternativas

Al tratarse de modelos de clasificación, se puede optar por utilizar regularización en la regresión logística para mejorar la generalización, o probar alternativas bayesianas, que permiten incorporar conocimiento previo y manejar la incertidumbre en las predicciones.

En este caso se utilizará regresión logística con regularización L1/L2, se aplicará calibración bayesiana de manera separada, y además se probará Naive Bayes como otra alternativa.

## 4.1 Comparación de Modelos Alternativos

Al evaluar los distintos modelos, se obtuvieron las siguientes métricas de desempeño, las cuales permiten comparar su capacidad de clasificación y generalización sobre los datos de validación.

Modelo	Accuracy	Precision	Recall	F1-Score
Logistic Reg	0.8791	0.8663	0.8791	0.8542
L1	0.8830	0.8726	0.8830	0.8596
L2	0.8839	0.8739	0.8839	0.8612
Naive Bayes	0.8646	0.8531	0.8646	0.8574
LogReg + Bayes Cal	0.8801	0.8701	0.8801	0.8534

Cuadro 4: Comparación de Métricas entre Modelos

La tabla muestra que todos los modelos presentan un desempeño alto y consistente, con accuracy alrededor de 0.87–0.88. La regularización L1/L2 mejora ligeramente las métricas de la regresión logística, mientras que Naive Bayes tiene un rendimiento un poco menor. La combinación de regresión logística con calibración bayesiana mantiene resultados similares, ofreciendo beneficios principalmente en la calibración de probabilidades.

### 4.1.1 validación Cruzada

Se utilizará validación cruzada para evaluar los modelos de manera más confiable, ya que permite estimar su desempeño promedio en distintos subconjuntos de los datos y reducir el riesgo de sobreajuste.

Modelo	Accuracy	Precision	Recall	F1-Score
L1	0.8727	0.8536	0.8727	0.8487
Logistic Reg + Cal Bayes	0.8706	0.8513	0.8706	0.8523
L2	0.8702	0.8492	0.8702	0.8448
Logistic Reg	0.8683	0.8458	0.8683	0.8425
Naive Bayes	0.8631	0.8512	0.8631	0.8557

Cuadro 5: Comparación de Métricas después de Validación Cruzada

La tabla muestra que los modelos con L1 y L2 tienen el mejor desempeño general, mientras que Naive Bayes mantiene un F1-Score competitivo pese a su menor Accuracy. Dado que nuestros datos están desbalanceados, resulta especialmente importante considerar métricas como Recall y F1-Score, que reflejan mejor el desempeño del modelo sobre la clase minoritaria.

## 5 Conclusiones

La replicación del estudio sobre predicción de hipertensión arterial mediante regresión logística permitió identificar aspectos críticos para su aplicación práctica:

- **Alta precisión global pero sensibilidad limitada:** Se obtuvo un accuracy del 87.9 %, sin embargo, la sensibilidad para detectar pacientes hipertensos fue muy baja (27-30 %), debido al desbalanceo de clases (solo 757 hipertensos de 5170 observaciones).
- **Modelos alternativos con mejoras marginales:** Los modelos con regularización L1/L2 mostraron ligeras mejoras en validación cruzada (Accuracy 87.3 %), mientras que Naive Bayes mantuvo un F1-Score competitivo. No obstante, ninguno resolvió satisfactoriamente el problema de baja sensibilidad.
- **Limitaciones principales:**
  - Desbalanceo de clases no tratado
  - Falta de transparencia en selección de variables del estudio original
  - Pérdida de datos por valores faltantes

En conclusión, aunque la replicación confirmó la viabilidad del enfoque, los modelos en su forma actual no son confiables para detección clínica debido a su incapacidad para identificar correctamente a la mayoría de pacientes hipertensos, destacando la necesidad de abordar el desbalanceo de datos con técnicas como SMOTE, oversampling, o ajuste de pesos de clase.

## Referencias

Tesillo Gómez, C. M., Escobar Arcaya, Y. A., & León Gutiérrez, E. D. (2021). Predicción de hipertensión arterial a través de un sistema de regresión logística [Recuperado el 10 de octubre de 2025]. *Innovación y Software*, 2(2), 60-74. <https://www.redalyc.org/articulo.oa?id=673870839005>