Facultad de Ciencias

Lo misional como guía de una facultad líder en el escenario de pospandemia
Equidad, transparencia y mérito

33° SIMPOSIO INTERNACIONAL DE ESTADÍSTICA 2024

CIENCIA DE DATOS

Universidad de Cartagena, Cartagena de Indias (Colombia).

30 de Julio al 02 de agosto de 2024.

https://simposioestadistica.unal.edu.co/

Departamento de Estadística
Facultad de Ciencias
Sede Bogotá

Apoya

ICETEX · Universidad Nacional Agraria La Molina · V CINESTAP

Organizan

Universidad de Cartagena · STEM+CO · Universidad Nacional de Colombia

UNIVERSIDAD
NACIONAL
DE COLOMBIA

PROYECTO **CULTURAL, CIENTÍFICO Y COLECTIVO** DE NACIÓN

Simposio

Estudio del Modelo de Regresión Logística y Aplicación en conductas relacionadas con Ideación Suicida

Christian Camilo Trilleras Mota. czambranot@unal.edu.co

Semillero de Estadística
Docente: Nubia Esteban Duarte. nesteband@unal.edu.co
Universidad Nacional de Colombia - Sede Manizales

Contenido

Introducción

La Regresión Logística es una técnica fundamental en Estadística y Ciencia de Datos, la cual se utiliza para modelar y predecir variables binarias y, en general, útil para resolver problemas de clasificación.

En este trabajo se presenta el desarrollo de los conceptos teóricos subyacentes a la regresión logística, destacando la importancia de esta metodología en el contexto de un problema real, asociado a un estudio de caso que involucra reportes de suicidios e ideación suicida, destacando la relevancia de esta técnica en la investigación y toma de decisiones.

Importancia de estudios sobre Ideación Suicida

- Realizar estudios sobre la ideación suicida es esencial para identificar factores de riesgo y desarrollar estrategias de prevención.
- Estos estudios ayudan a mejorar los servicios de salud mental, capacitar a profesionales y formular políticas públicas efectivas.
- Además, proporcionan conocimiento para crear tratamientos personalizados y reducir el impacto económico y social del suicidio.

En conjunto, estas investigaciones son clave para abordar una problemática de salud pública y mejorar el bienestar general de la sociedad.

Banco de datos

- La base de datos utilizada corresponde a una recopilación realizada en la ciudad de Bogotá.
- Son individuos relacionados con la conducta suicida (pacientes con presencia de dicha conducta), suministrada por la Secretaría Distrital de Salud.
- El sistema de vigilancia epidemiológica de la conducta suicida, conocido como **SISVECOS**, recopila datos relativos a casos de ideación, amenaza e intento de suicidio.

EDAD	CLASIFICACION_DE_LA_CONDUCTA	SEXO	INTENTOS_PREVIOS	PSIQUIATRIA	CONSUMADO
85	amenaza suicida	Mujer	NO	SI	0
75	ideacion suicida	Mujer	SI	SI	0
47	ideacion suicida	Mujer	NO	SI	0
75	ideacion suicida	Hombre	NO	NO	0
		...			
29	suicidio consumado	Hombre	NO	NO	1

Diccionario de las variables

- **Suicidio Consumado:** Se trata de un acto autolesivo intencionado que resulta en la muerte.
- **Sexo:** Clasifica a los individuos en dos categorías distintas, **Hombre** y **Mujer**.
- **Intentos Previos:** Clasifica a los individuos en dos categorías distintas, **Sí** y **No**. Indica la presencia o ausencia de intentos previos de suicidio.
- **Edad:** Registra la edad de los individuos en años, abarcando un rango que va desde los 8 hasta los 94 años.
- **Psiquiatría:** Clasifica a los individuos en dos categorías distintas. **Sí**, indica que la persona fue remitida al servicio de salud de psiquiatría a partir de la notificación de casos de conducta suicida y **No**, indica el caso contrario.

Contenido

Planteamiento del problema

Teoría	Estudio de caso	Ilustración
Comprender la regresión logística y demostrar su utilidad práctica en la investigación y la toma de decisiones.	Se presenta un estudio de caso que aborda reportes de suicidios y conductas suicidas en Bogotá.	Se ilustra cómo la regresión logística puede ser una herramienta valiosa para analizar y predecir eventos críticos en la sociedad, proporcionando información significativa para la intervención y prevención en temas de salud pública

Contenido

Justificación

- La Regresión Logística es crucial en situaciones donde se necesita predecir la probabilidad de un resultado binario, **por ejemplo la probabilidad de padecer, o no, una determinada enfermedad.**
- En muchos casos, puede ofrecer un buen rendimiento predictivo en comparación con métodos más complejos, especialmente cuando se enfrenta a problemas de clasificación binaria.
- Su versatilidad hace que sea una herramienta valiosa en el análisis estadístico en muchas disciplinas, como:
  Investigaciones clínicas, Epidemiológicas, Médicas, entre otras.

Contenido

Objetivos

Objetivo general

Exponer la teoría asociada a Modelos de Regresión Logística con aplicaciones en datos reales.

Objetivos especificos

1. Presentar la fundamentación teórica y práctica de los modelos de regresión logística con aplicaciones en datos reales.
2. Utilizar el Programa R y el lenguaje de programación Python para realizar las aplicaciones que ilustran la teoría asociada a Regresión Logística, en escenarios de interés práctico.

Contenido

Metodología

- Preparación del diccionario de variables.
- Análisis descriptivo de las variables: Edad, Conducta, Sexo, Intentos Previos, Psiquiatría.
- Algunos análisis conjuntos de las variables para ver como se complementan.
- Posteriormente se ajustaron los modelos de Regresión Logística.

La preparación de los datos, análisis y ajuste del modelo fue realizado utilizando los lenguajes de programación Python y RStudio.

Contenido

Referente teórico

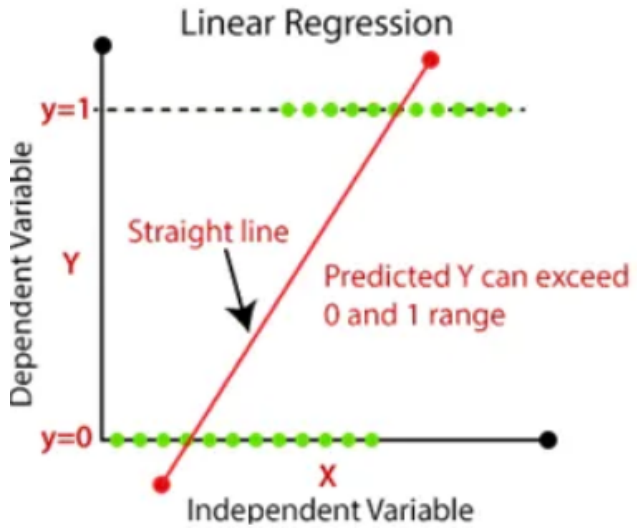
Modelo de Probabilidad Lineal (MPL)

- Sea $y_i \sim \text{Bernoulli}(p_i)$, donde $p_i = P(y_i = 1 \mid \mathbf{x})$.
- La forma especifica del modelo es,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

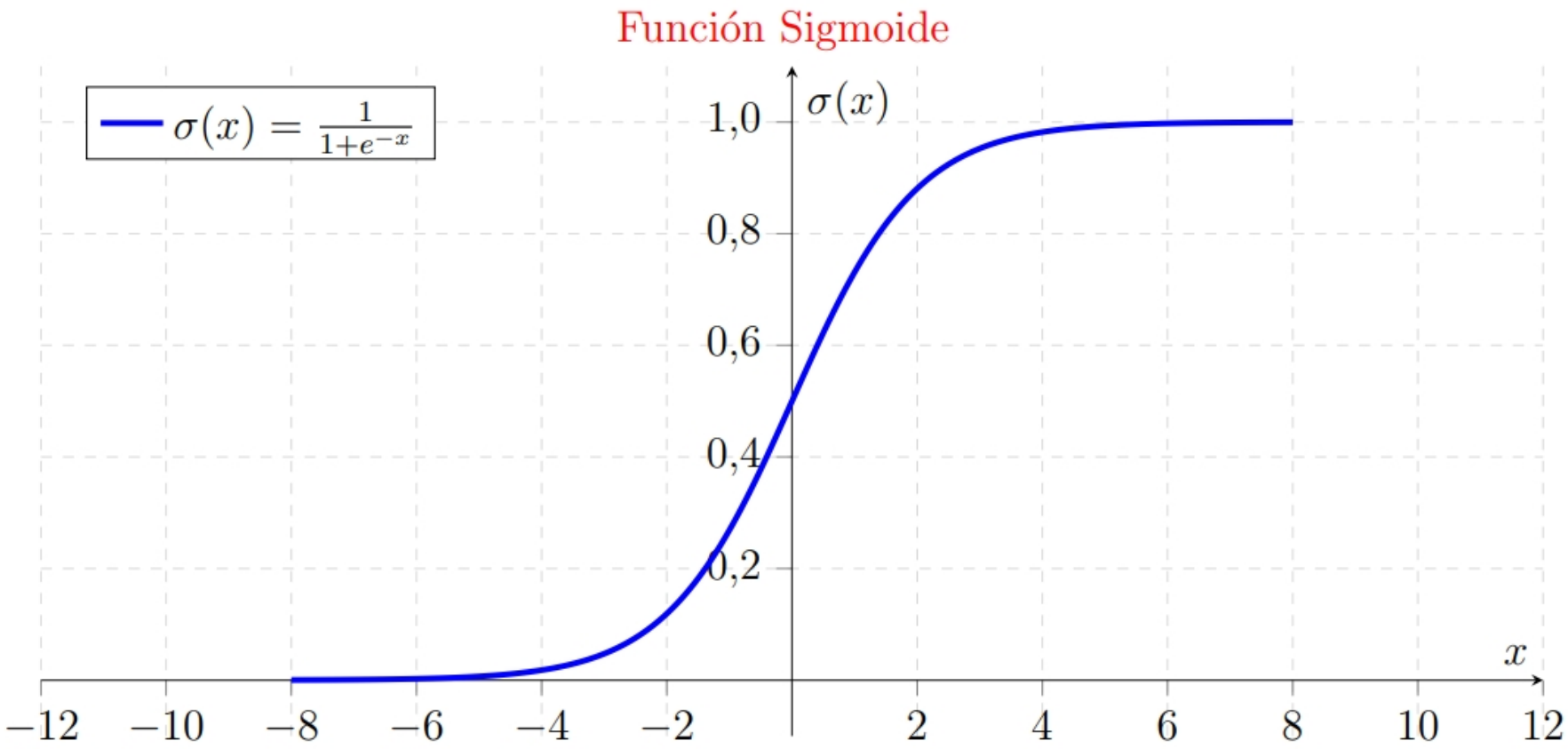
tal que,

$$E(y_i) = p_i = P(y_i = 1 \mid \mathbf{X}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$$



Referente teórico

Función Sigmoide

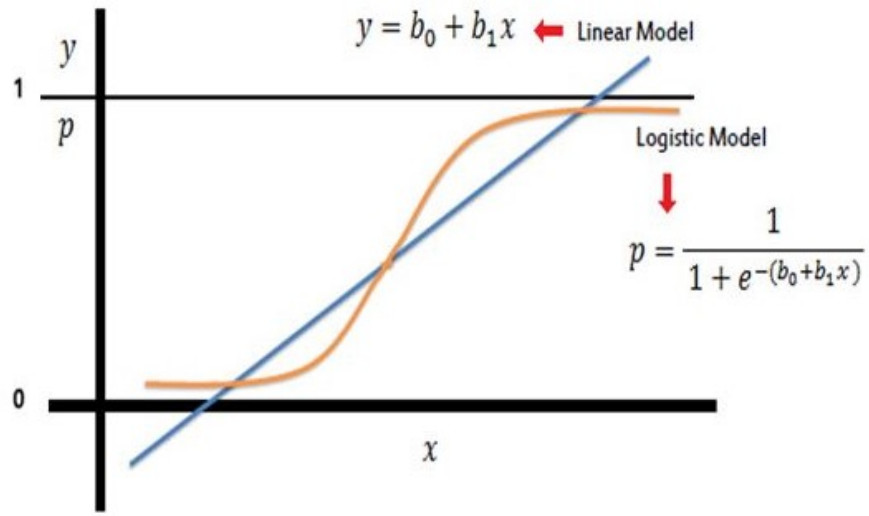


Referente teórico

Modelo de Regresión Logística

El modelo de regresión logística está dado por:

$$\pi_i = E(y_i) = P(y_i = 1 \mid \mathbf{X}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki})}}$$



Referente teórico

Transformación Logit

La función logit está dada por:

$$\text{logit}(x) = \ln\left(\frac{x}{1-x}\right)$$

Por lo tanto, para nuestro caso en particular:

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} = \boldsymbol{\beta}^T \mathbf{x}_i$$

Donde,

$$\frac{P(y_i = 1 \mid \mathbf{X})}{P(y_i = 0 \mid \mathbf{X})} = \frac{\pi_i}{1-\pi_i} = e^{\boldsymbol{\beta}^T \mathbf{x}_i}$$

se denomina **Odds**.

Referente teórico

- Para una variable x_{ji}, al incrementar x_{ji} en una unidad, las odds cambian de:

$$\frac{\pi_i}{1-\pi_i} = e^{\boldsymbol{\beta}^T \mathbf{x}_i}$$

a

$$\frac{\pi'_i}{1-\pi'_i} = e^{\beta_0+\beta_1 x_{1i}+\cdots+\beta_j(x_{ji}+1)+\cdots+\beta_k x_{ki}}$$

- La razón de odds (odds ratio) es:

$$\text{OR} = \frac{\frac{\pi'_i}{1-\pi'_i}}{\frac{\pi_i}{1-\pi_i}} = \frac{e^{\beta_0+\beta_1 x_{1i}+\cdots+\beta_j(x_{ji}+1)+\cdots+\beta_k x_{ki}}}{e^{\beta_0+\beta_1 x_{1i}+\cdots+\beta_j x_{ji}+\cdots+\beta_k x_{ki}}}$$

- Simplificando, obtenemos:

$$\text{OR} = e^{\beta_j}$$

- Por lo tanto, e^{β_j} representa el cambio en las odds de π_i asociado a un incremento unitario en x_{ji}.

Referente teórico

Odds Ratio (OR)

$$OR_j = e^{\beta_j}$$

- **OR** $= 1$: Sin efecto en odds del evento de interés.
- **OR** > 1: Incremento en odds del evento de interés.
- **OR** < 1: Decremento en odds del evento de interés.

Ejemplo:

$$\beta_2 = 0.5 \implies OR_2 = e^{0.5} \approx 1.65$$

Un incremento unitario en X_2 aumenta las odds en 65%.

Contenido

Resultados

Análisis descriptivos

Al analizar las variables de interés individualmente, se observa que presentan las siguientes características descriptivas:

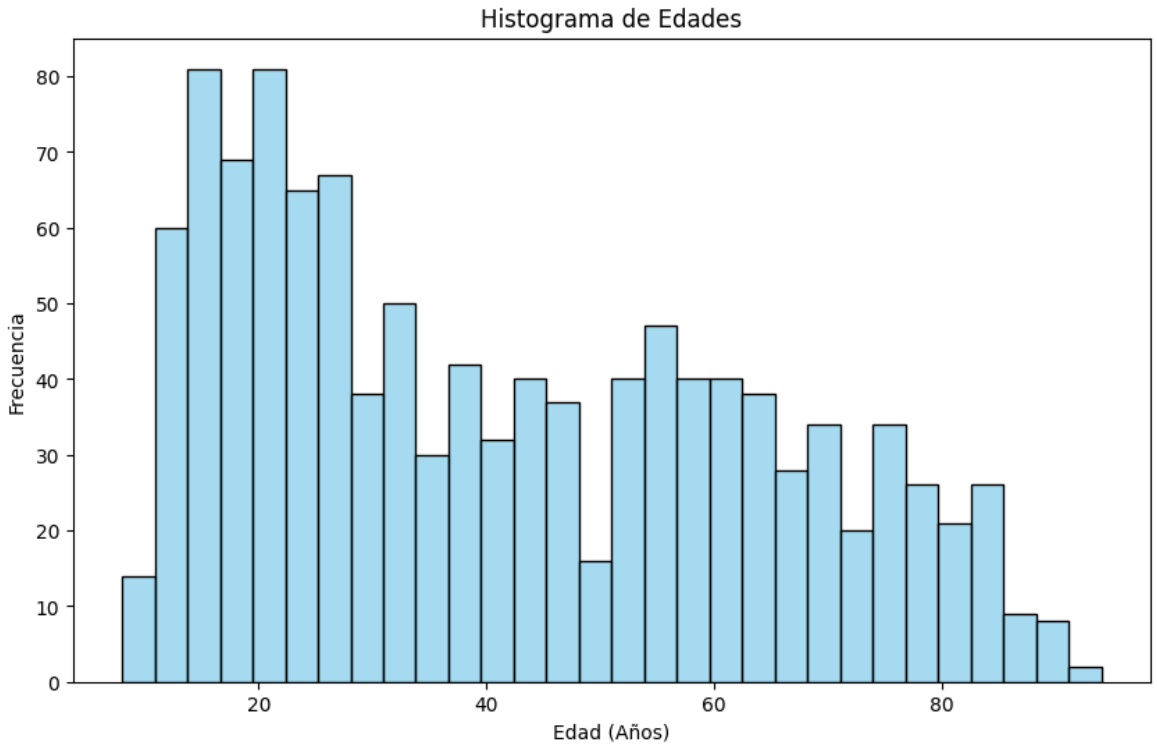


Figura: Histograma de frecuencias de la variable **EDAD**

Resultados

Análisis descriptivos

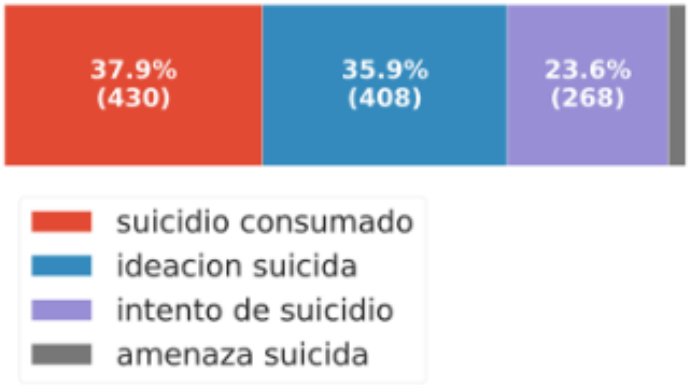


Figura: Frecuencias de la variable **Clasificación de la conducta**

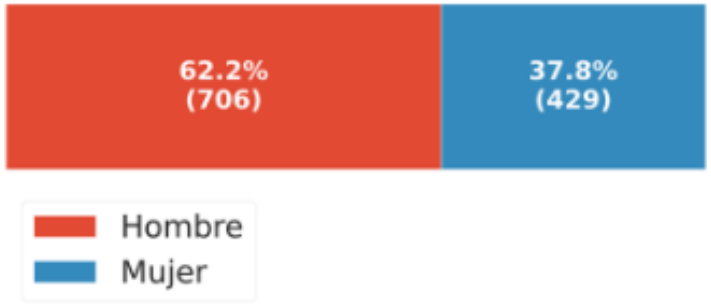


Figura: Frecuencias de la variable **Sexo**

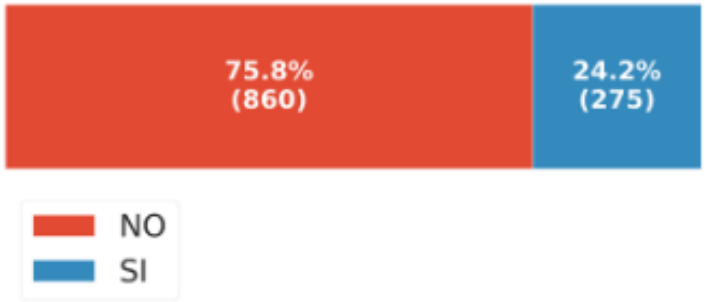


Figura: Frecuencias de la variable **Intentos previos**

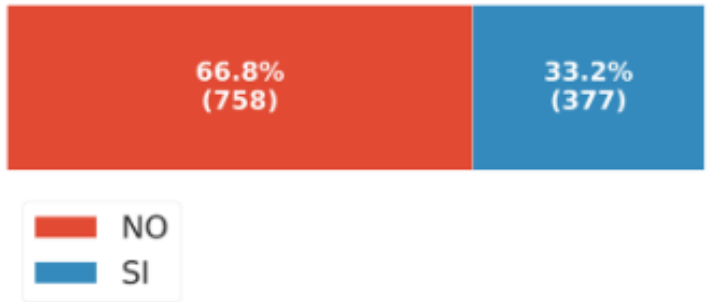


Figura: Frecuencias de la variable **Psiquiatría**

Resultados

Al analizar las variables de interés conjuntamente, se observa que presentan las siguientes características descriptivas:

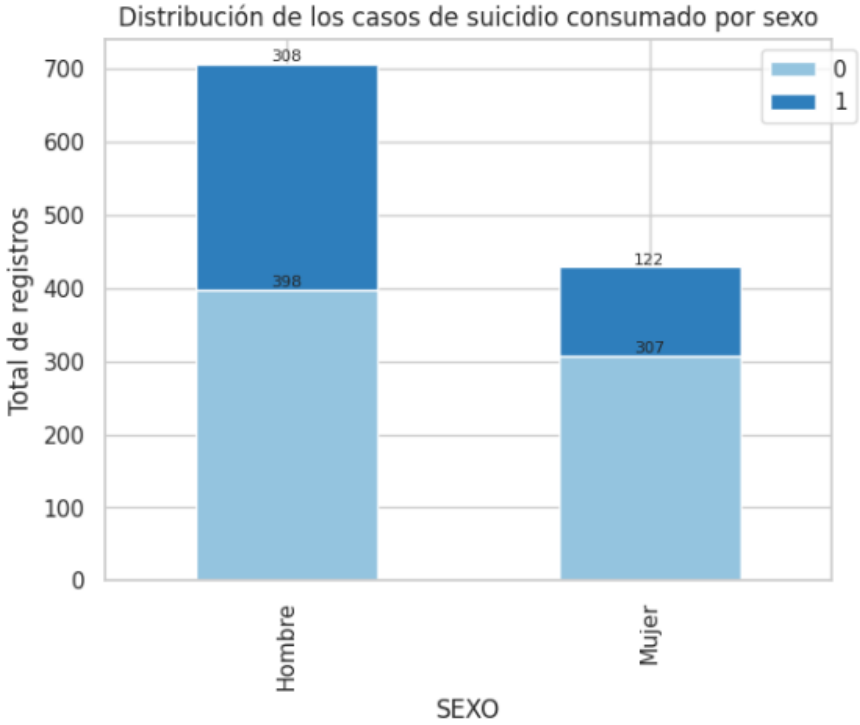


Figura: Distribución de los casos de suicidio consumado por sexo

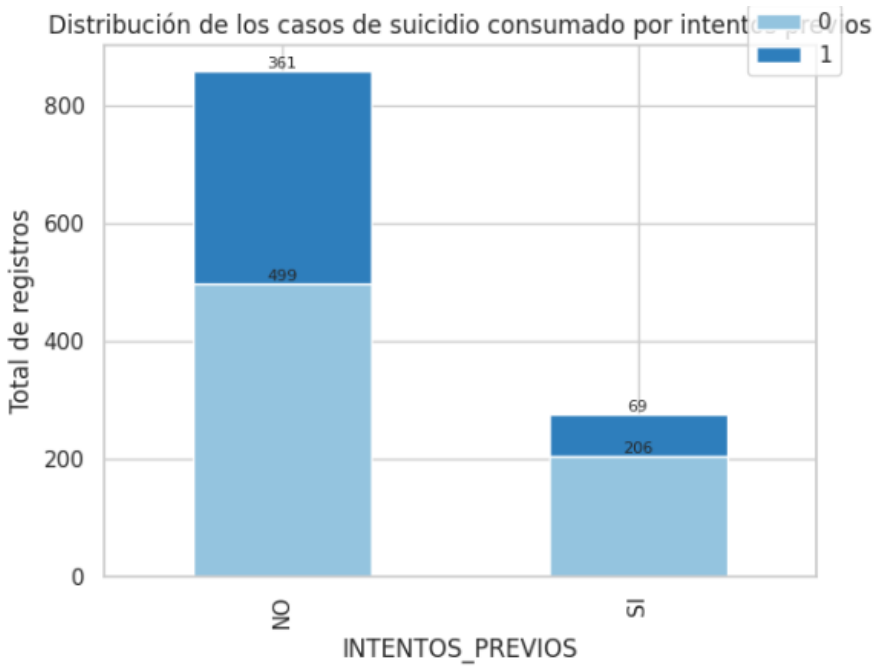


Figura: Distribución de los casos de suicidio consumado por intentos previos

Resultados

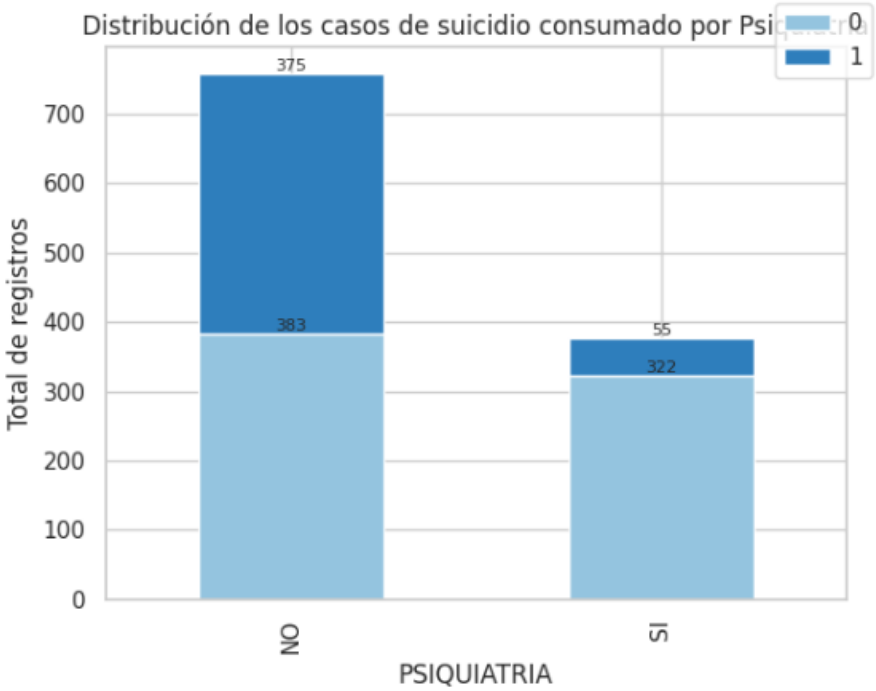


Figura: Distribución de los casos de suicidio consumado por psiquiatría

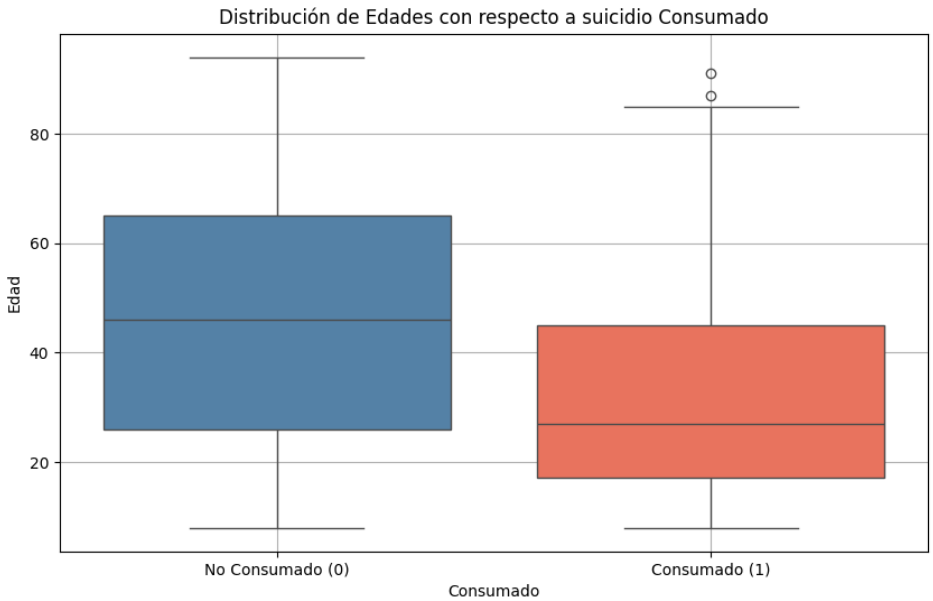


Figura: Distribución de la edad con respecto a suicidio consumado

Resultados

Modelos de Regresión Logística

Consideraremos dos modelos, uno en el cual la variable de respuesta Consumado se predice en función de la Edad del sujeto, y otro en función de las variables Sexo, Psiquiatría, Intentos Previos y Edad.

Modelo 1:

$$\text{Logit}(\pi) = \beta_0 + \beta_1 \cdot \text{Edad}$$

Modelo 2:

$$\text{Logit}(\pi) = \beta_0 + \beta_1 \cdot \text{Sexo\_Mujer} + \beta_2 \cdot \text{Edad} + \beta_3 \cdot \text{Intentos\_Previos\_Si} + \beta_4 \cdot \text{Psiquiatría\_Si}$$

Resultados

Resultados para el modelo de Regresión Logística (Modelo 1)

A continuación, se presentan los resultados obtenidos relacionados con la estimación de los parámetros del modelo.

Tabla: Coeficientes de regresión logística estimados.

Parámetro	Estimación	Std. Error	Valor Z	P- valor
β_0	0.68425	0.14850	4.61453	5.89×10^{-6}
β_1	-0.029887	0.00448	-8.48864	$< 2 \times 10^{-16}$

Tabla: Intervalos de confianza para los Odds Ratio de β_0 y β_1.

Parámetro	OR	OR 25%	OR 97.5%	P-valor
β_0	1.9821	1.4823	2.6505	3.940854×10^{-6}
β_1	0.9706	0.9640	0.9773	2.105584×10^{-17}

Resultados

En resumen tenemos:

- Las estimaciones de máxima verosimilitud son:

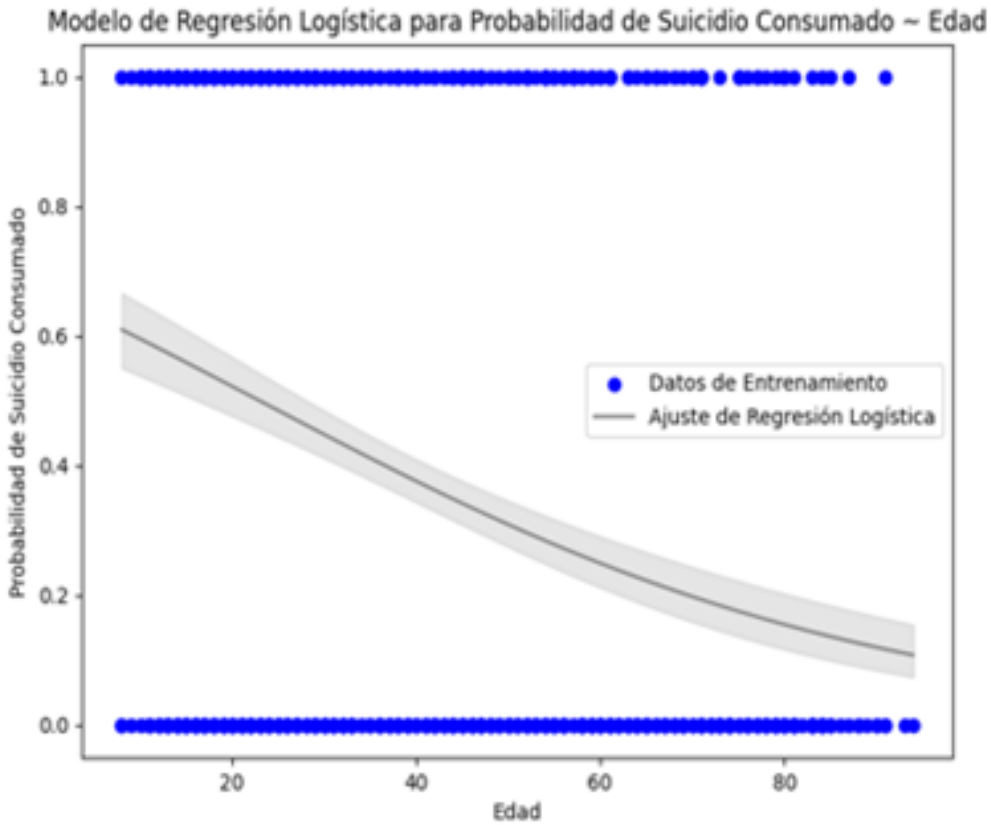
$\widehat{\beta_0}$= 0,68425 y $\widehat{\beta_1}$= −0,029887

- Los valores ajustados vienen dados por la ecuación:

$$\hat{\pi}(Edad) = \frac{e^{0{,}68425 - 0{,}029887 \cdot \text{Edad}}}{1 + e^{0{,}68425 - 0{,}029887 \cdot \text{Edad}}}$$

- El logit estimado viene dado por la ecuación:

$$\widehat{Logit}(\pi) = 0{,}68425 - 0{,}029887 \cdot \text{Edad}$$



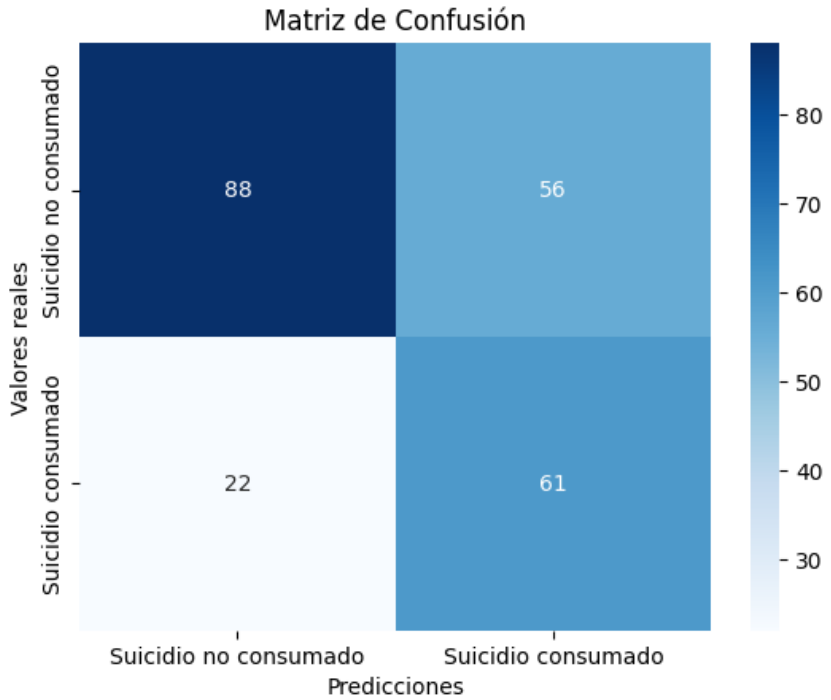
Resultados

Evaluación del modelo

Matriz de confusión

Matriz de Confusión		
	Predicción No Consumado	**Predicción Consumado**
Verdadero No Consumado	TN	FP
Verdadero Consumado	FN	TP

En particular, para nuestro modelo tenemos:



Resultados

De lo anterior, el modelo es capaz de clasificar correctamente el 66% de las observaciones cuando se emplean los datos de prueba. Este porcentaje de clasificación se denomina **Accuracy** y se obtiene de la siguiente manera:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Resultados

Evaluación mediante Curva ROC y AUC

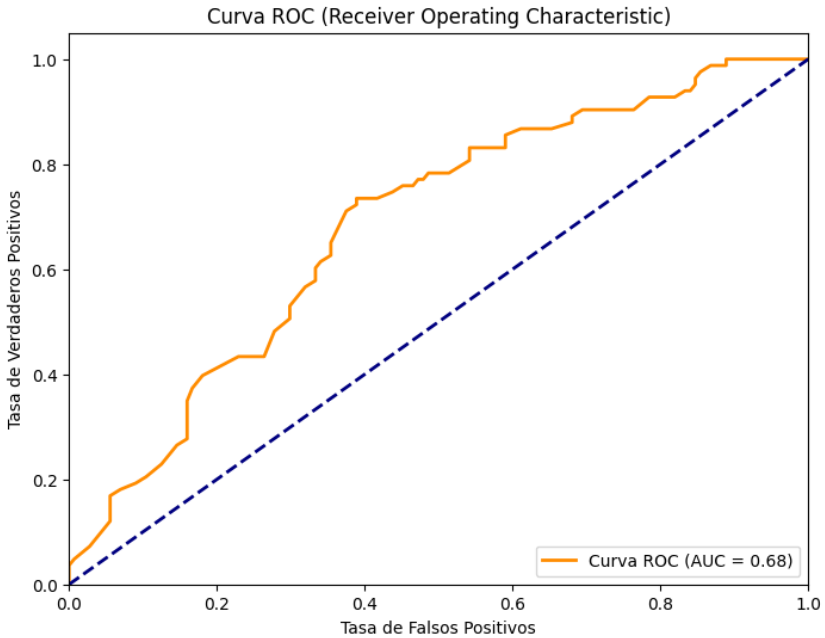


Figura: Curva ROC para el modelo de Regresión Logística Simple

Donde,

$$\text{FPR} = \frac{FP}{FP + TN} \quad \text{y} \quad \text{TPR} = \frac{TP}{TP + FN}$$

Resultados

Resultados para el modelo de Regresión Logística (Modelo 2)

A continuación, se presentan los resultados obtenidos relacionados con la estimación de los parámetros del modelo.

Parámetro	OR	OR 25%	OR 97.5%	P-valor
β_0	5.3980	3.6944	7.8871	2.919910×10^{-18}
β_1	0.4757	0.3446	0.6569	6.390266×10^{-6}
β_2	0.9671	0.9600	0.9744	1.558229×10^{-18}
β_3	0.4877	0.3304	0.7199	3.018808×10^{-4}
β_4	0.1961	0.1359	0.2849	1.177605×10^{-17}

Tabla: Intervalos de confianza para los Odds Ratio de los coeficientes del modelo.

Resultados

Parámetro	Estimación	Std. Error	Valor Z	P- valor
β_0	1.68746	0.19394	8.715	$< 2 \times 10^{-16}$
β_1	-0.74247	0.16548	-4.510	6.53×10^{-06}
β_2	-0.033492	0.004	-8.785	2×10^{-16}
β_3	-0.71816	0.19441	-3.614	0.000596
β_4	-1.6290	0.192	-8.555	$< 2 \times 10^{-16}$

Tabla: Coeficientes de regresión logística múltiple estimados.

- **El Logit estimado viene dado por la ecuación:**

$$\hat{\text{Logit}}(\pi) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Sexo\_Mujer} + \hat{\beta}_2 \cdot \text{Edad} + \hat{\beta}_3 \cdot \text{Intentos\_Previos\_Si} + \hat{\beta}_4 \cdot \text{Psiquiatría\_Si}$$

Donde cada $\hat{\beta}_i$ con $i = 0, \cdots, 4$ corresponde a los valores estimados en la tabla.

Resultados

Evaluación del modelo

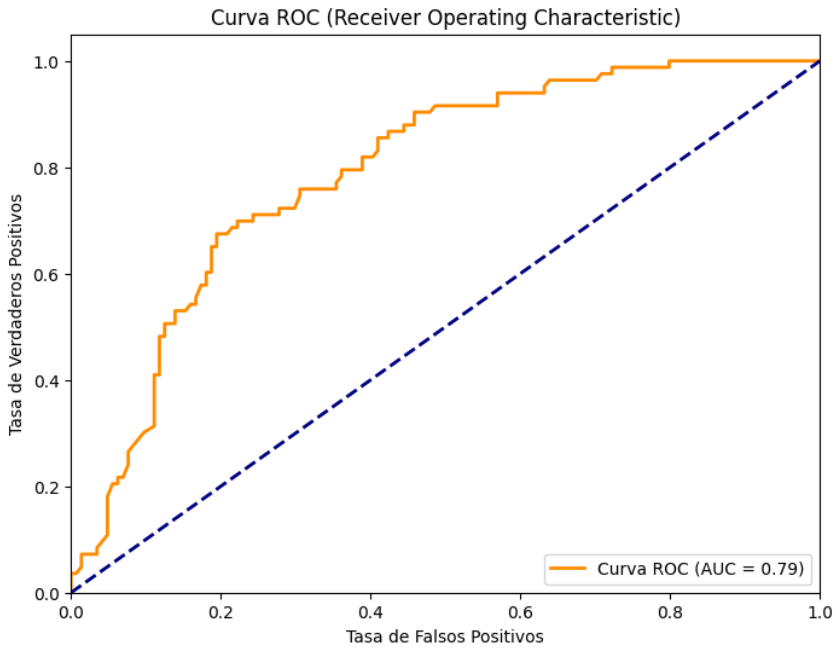


Figura: Curva ROC

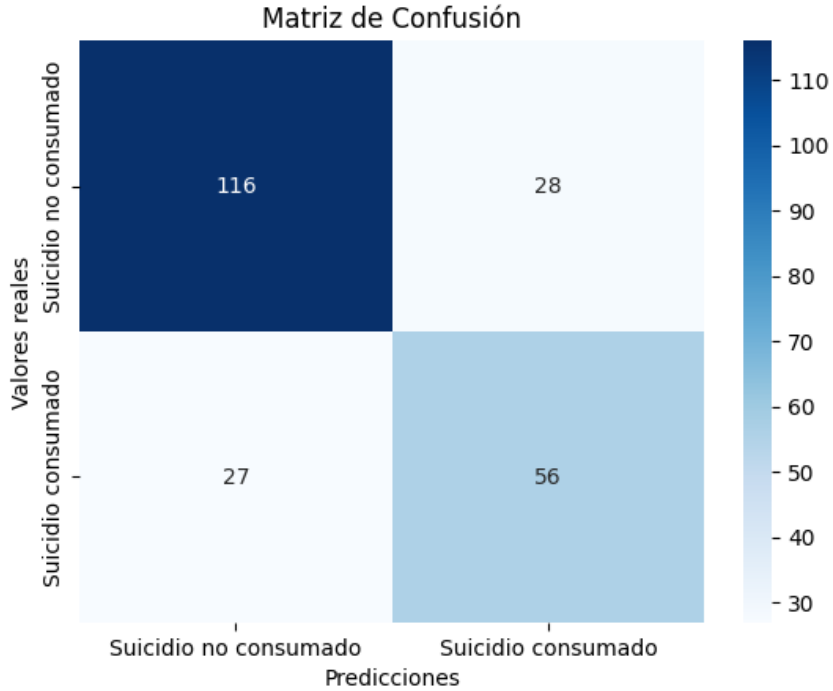


Figura: Matriz de confusión

Resultados

El modelo es capaz de clasificar correctamente el 76%. Además, usando variables como edad, sexo, antecedentes psiquiátricos y previos intentos de suicidio, muestra una notable capacidad para predecir suicidios consumados, con un AUC del 79%.

Contenido

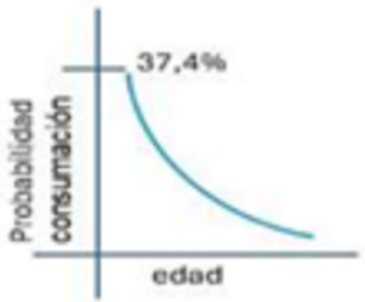
Discusión de resultados

Prevalencia del fenómeno

Según el género

Significativamente mayor en hombres

Según la edad



Significativamente mayor en individuos jóvenes

De la muestra estudiada el **37,4%** fueron **suicidos consumados**

El riesgo se reduce con la edad

Se resalta que la inclusión de variables en el modelo, como Psiquiatría, Sexo, Intentos Previos, además de Edad, proporcionan una mejor comprensión de los factores asociados al riesgo de suicidio consumado.

Contenido

Conclusiones

El estudio es de gran relevancia, pues consigue conjugar conceptos matemáticos, probabilísticos y de estadística, asociados a la Regresión Logística, que se aplican de una forma clara a un problema real y reciente que es de interés social.

Tras probar varios modelos de clasificación con validación cruzada y ajustar los hiperparámetros, se confirmó que el modelo de regresión logística es el más adecuado, consolidándose como la mejor opción para resolver el problema de clasificación.

Con este estudio, se espera contribuir a la investigación del suicidio, un problema social que requiere apoyo gubernamental para campañas efectivas de prevención y tratamiento.

Referencias Bibliográficas


[Abbott, 1985] Abbott, R. D. (1985). Logistic regression in survival analysis. American journal of epidemiology, 121(3):465–471.


[Andreas C. Müller, 2016] Andreas C. Müller, S. G. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media, 1 edition.


[Bishop, 2006] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Information science and statistics. Springer, 1st ed. 2006. corr. 2nd printing edition.


[Bernal, James Lopez, Steven Cummins y Antonio Gasparrini, 2017] "Interrupted time series regression for the evaluation of public health interventions: a tutorial". En: International journal of epidemiology 46.1, p´ags. 348-355.


[George Casella, 2001] George Casella, R. L. B. (2001). Statistical Inference. Duxbury Press, 2 edition.


[Hosmer Jr et al., 2013] Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). Applied logistic regression, volume 398. John Wiley Sons.


[Montgomery et al., 2002] Montgomery, D. C., Peck, E. A., Vining, G. G., et al. (2002). Introducción al análisis de regresión lineal.

GRACIAS