XVII ENCUENTRO DEPARTAMENTAL DE SEMILLEROS DE INVESTIGACIÓN
RREDSI NODO CALDAS

II ENCUENTRO DEPARTAMENTAL DE SEMILLEROS DE INVESTIGACIÓN
REDCOLSI NODO CALDAS

PROYECTO TERMINADO

ESTUDIO DEL MODELO DE REGRESIÓN LOGÍSTICA Y SUS APLICACIONES

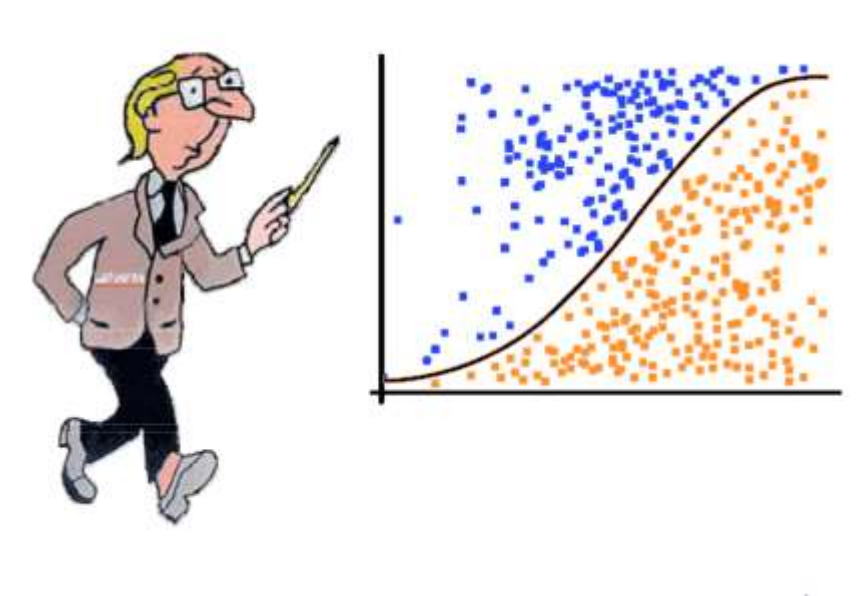
Semillero de Estadística

Christian Camilo Trilleras Mota. czambranot@unal.edu.co
Universidad Nacional de Colombia - Sede Manizales

CONTENIDO

INTRODUCCIÓN

Regresión Logística – Ejemplos



Estado del tiempo (seco/lluvioso).

Resultado de una prueba diagnóstica (+/-).

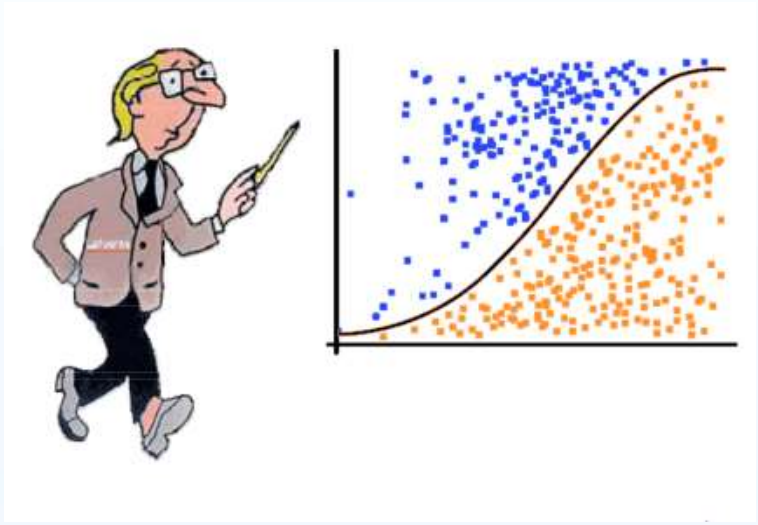
Decisión en relación a un producto ofrecido (compra/no compra).

Resultado de un examen de matemáticas (aprueba/no aprueba).

Resultado del Índice de Masa Corporal IMC (sobrepeso/Sin sobrepeso).

INTRODUCCIÓN

Regresión Logística – Ejemplos prácticos



- En un servicio hospitalario de quemados se quiere construir Un modelo predictivo **para la muerte de los pacientes que ingresan**. (Asociar con la edad, % de quemadura y otras variables).

- Evaluar la influencia de un régimen (**Convencional o Experimental**) de atención de cuidados de enfermería sobre la recuperación de los pacientes.

- Dado un estudiante que entra a la universidad con un perfil conocido, ¿cuál es la probabilidad de que **abandone sus estudios**?

INTRODUCCIÓN

- La regresión logística es una herramienta estadística fundamental en el campo de la ciencia de datos y el análisis predictivo.

- Esta metodología es esencial para comprender y predecir resultados binarios o categóricos en diversas áreas.

- Permite modelar relaciones entre una variable binaria (dependiente) y una o más variables independientes (predictoras), que pueden ser numéricas o categóricas.

Antes de intentar resolver un problema práctico, es crucial entender la fundamentación teórica. Esto no solo permite comprender la metodología de la regresión logística como un método específico, sino que también facilita la aplicación de conocimientos en diversos campos, respaldando la toma de decisiones informadas y efectivas

INTRODUCCIÓN

Información de los datos para la aplicación

- La base de datos utilizada corresponde a una recopilación realizada en la ciudad de Bogotá. 2021.
- 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
0	85	amenaza suicida	Mujer	NO	SI	0
1	75	ideacion suicida	Mujer	SI	SI	0
2	47	ideacion suicida	Mujer	NO	SI	0
3	75	ideacion suicida	Hombre	NO	NO	0
4	76	ideacion suicida	Hombre	NO	SI	0
...	...	...	...	...	...	...
1130	29	suicidio consumado	Hombre	NO	NO	1

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

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 método 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,**
- **Psicosociales,**
- **Sociales y**
- **Financieras, entre otras.**

OBJETIVOS

Objetivo general

Presentar y formalizar la teoría asociada a Modelos de Regresión Logística Simple con aplicaciones en datos reales.

Objetivos específicos

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.

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 varios modelos de Regresión Logística.**

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

REFERENTE TEÓRICO

- **Distribución Bernoulli**

Considérese un experimento aleatorio que solo puede dar lugar a dos resultados posibles mutuamente excluyentes, llamados “éxito” y “fracaso”. Sea p la probabilidad de éxito, por lo que la probabilidad de fracaso es 1- p.
Sea X variable aleatoria tal que, si P(X = 1) = p y P(X = 0) = 1 – p, entonces se dice que X tiene distribución de Bernoulli de parámetro p y su función de densidad está dada por:

$$P(X = x) = f(x) = \begin{cases} p^x \cdot (1 - \mathrm{p})^{1-x} \text{ si x } = 0,1 \\ 0 \text{ en otro caso} \end{cases}$$

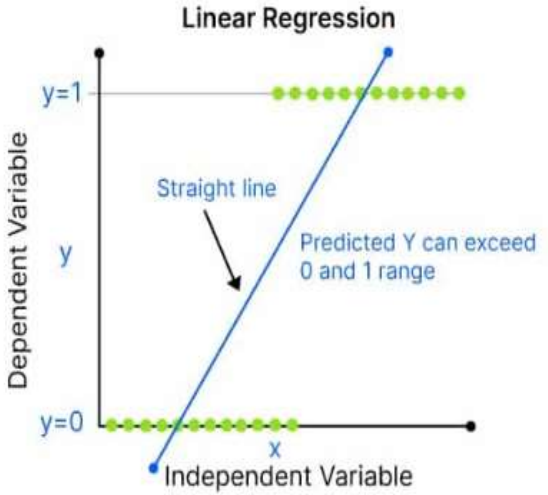
Propiedades

$\mu = E[x] = p$; $\sigma^2 = \mathrm{Var}(X) = E[(X - \mu)^2] = p(1- p)$

REFERENTE TEÓRICO

- **Modelo de probabilidad lineal**

El Modelo de Probabilidad Lineal (MPL) es una técnica simple y directa para predecir probabilidades utilizando una relación lineal entre una variable dependiente binaria y una o más variables independientes.



La forma especifica del modelo es,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Tal que satisface

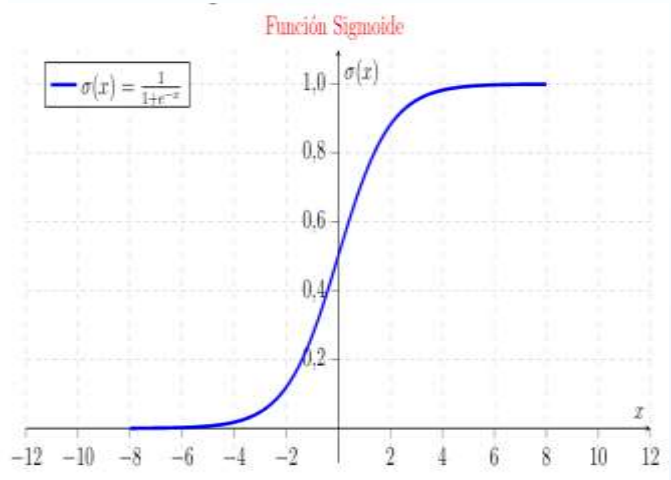
$$E[Y_i \mid x_i] = P(Y_i = 1 \mid x_i) = \beta_0 + \beta_1 x_i$$

Dado que se trata de una función lineal, la probabilidad puede aumentar de manera lineal en algunas ocasiones. Por lo tanto, no existe una garantía absoluta de que siempre se cumpla la condición $0 \leq E[Y_i \mid x_i] \leq 1$.

REFERENTE TEÓRICO

- **Función Sigmoide**

La función sigmoide se utiliza en la regresión logística para Transformar una combinación lineal de variables independientes en una probabilidad que varía entre 0 y 1. Esta transformación garantiza que las predicciones sean válidas probabilidades y captura la relación no lineal entre las variables, facilitando la interpretación de los resultados del modelo.



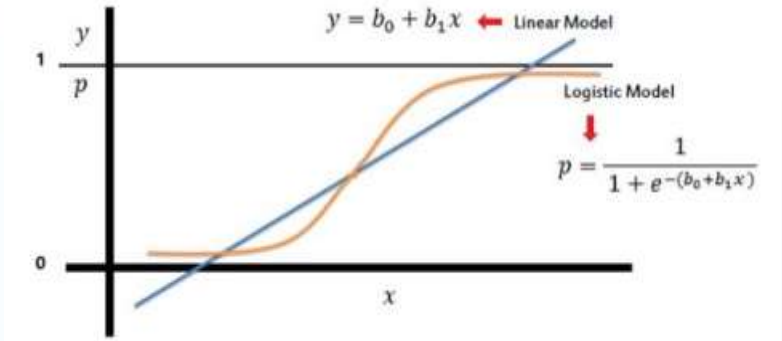
La función sigmoide se expresa matemáticamente como:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

REFERENTE TEÓRICO

- **Forma Específica del Modelo de Regresión Logística**

Consideremos $\pi(x) = E[Y \mid x]$, que representa la media condicional de Y dado x tal que para modelar la relación entre $\pi(x)$ y x en el contexto de la regresión logística, Utilizamos la siguiente expresión:



$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Vale la pena señalar que esta función es una instancia de la función sigmoide, que se caracteriza por su forma particular enunciada anteriormente y su capacidad para transformar una amplia gama de valores en un determinado rango.

REFERENTE TEÓRICO

- **Transformación Logit**

La transformación logit, denotada como Logit($\pi(x)$) o simplemente g(x), es una parte central de la regresión logística. Esta transformación se define como:

$$\mathbf{Logit}(\boldsymbol{\pi(x)}) = \boldsymbol{g(x)} = \ln\left(\frac{\boldsymbol{\pi(x)}}{\mathbf{1}-\boldsymbol{\pi(x)}}\right) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_i}$$

La transformación logit convierte una probabilidad en una escala logarítmica lineal, permitiendo modelar relaciones no lineales de manera lineal.

Observación: La expresión $\frac{\boldsymbol{\pi(x)}}{\mathbf{1}-\boldsymbol{\pi(x)}}$ se denomina Odds y representa la relación entre la probabilidad de que ocurra un evento ($\pi(x)$) y la probabilidad de que no ocurra ($1-\pi(x)$) y se interpreta como la tasa de fallos con respecto a la de aciertos.

REFERENTE TEÓRICO

- **Modelo de regresión Logística**

Sea Y = E[Y | x] + ε un modelo de regresión logística, donde E[y/x] = $\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1+e^{\beta_0 + \beta_1 x}}$.

- **Odds Ratio (OR)**

La razón de probabilidades (Odds Ratio (OR)) es una medida que compara las probabilidades de éxito entre dos grupos (uno con la variable predictora y otro sin ella) y se usa para evaluar cómo la presencia o ausencia de la variable predictora afecta las probabilidades del evento de interés. Dicha medida está dada por:

$$\psi = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} = e^{\beta_1}$$

La interpretación de la Odds Ratio en regresión logística depende de los valores que puede tomar, por ejemplo:

REFERENTE TEÓRICO

- **OR = 1:** Cuando la Odds Ratio es igual a 1, esto sugiere que no hay efecto de la variable predictora en las probabilidades del evento. En otras palabras, la variable no tiene un impacto en el evento de interés.

- **OR > 1:** Si la Odds Ratio es mayor que 1, esto significa que la variable predictora en cuestión está asociada con un aumento en las probabilidades del evento de interés. Es decir, los odds son aproximadamente $(e^{\beta_1} - 1) \cdot 100\%$ más altos si x=1 que si x=0

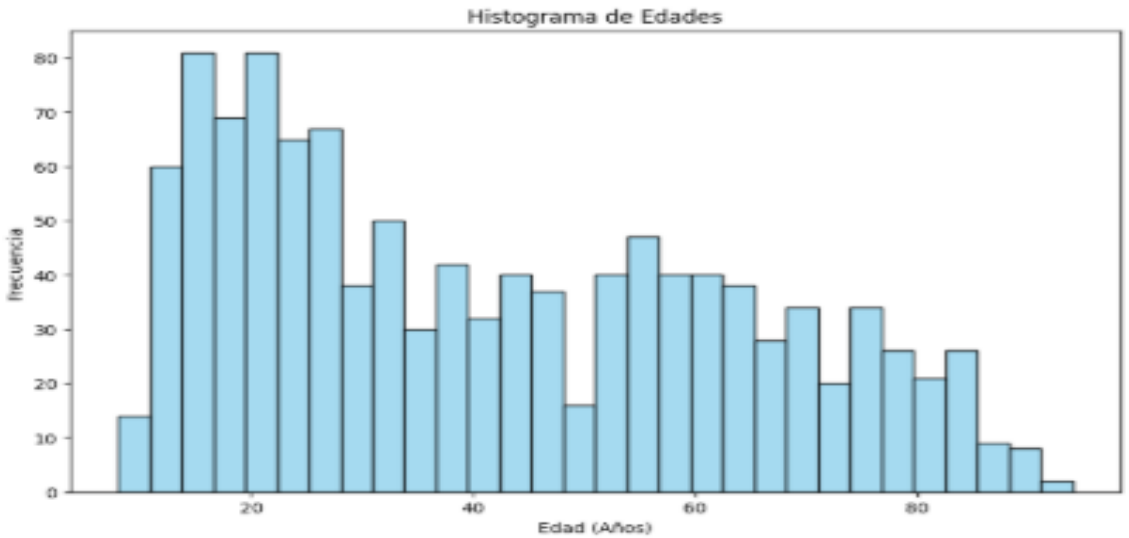
- **OR <1:** Si la Odds Ratio es menor que 1, esto indica que la variable predictora está asociada con una disminución en las probabilidades del evento de interés. Es decir, los odds son aproximadamente $(1 - e^{\beta_1}) \cdot 100\%$ más bajos si x=1 que si x=0.

RESULTADOS

- **Análisis descriptivos**

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

Histograma de frecuencias de la variable **EDAD**



RESULTADOS

- **Análisis descriptivos**

Frecuencias variable **Clasificación de la Conducta** y variable **Sexo**

▪ Variable Clasificación de la conducta

▪ Variable Sexo

Frecuencias variable **Intentos previos** y variable **Psiquiatría**

• Variable Intentos previos

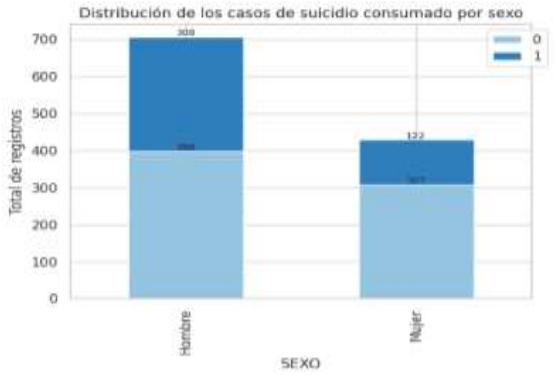
• Variable Psiquiatría

RESULTADOS

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

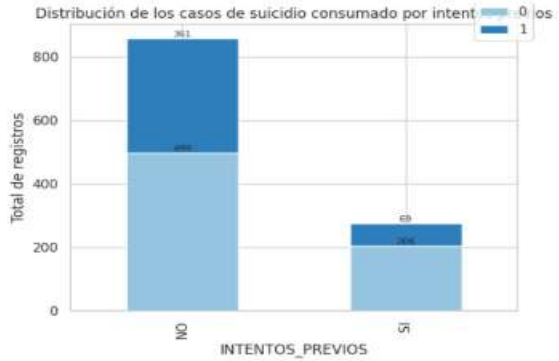
Distribución de los casos de **suicidio consumado** por **sexo**

- Variable consumado vs Variable sexo



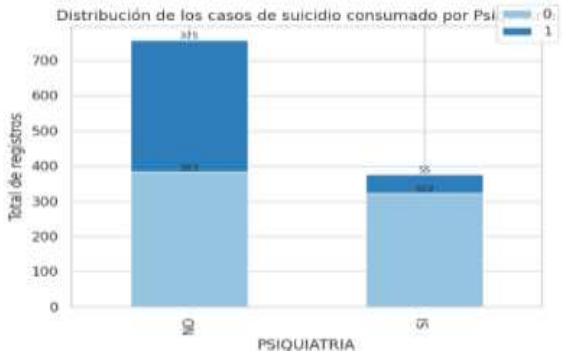
Distribución de los casos de **suicidio consumado** por **intentos previos**

- Variable consumado vs Variable intentos previos



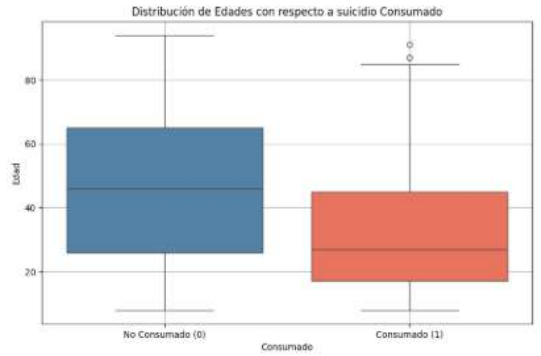
Distribución de los casos de **suicidio consumado** por **psiquiatría**

- Variable consumado vs Variable psiquiatría



Distribución de los casos de suicidio consumado por **sexo**

- Variable consumado vs Variable edad



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**. Desde la perspectiva matemática, podemos expresar los modelos de la siguiente manera.

Modelo 1:

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

Modelo 2:

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

Donde:

- β_0 : Corresponde al valor esperado del logaritmo de odds cuando todos los predictores son cero
- β_i : Son los coeficientes de regresión parcial de cada predictor e indican el cambio promedio del logaritmo de odds al incrementar en una unidad la variable predictora x_i , manteniéndose constantes el resto de variables.

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 1: Coeficientes de regresión logística estimados

Parámetro	Estimación	Std. Error	Valor Z	P- valor
β_0	0.68425	0.14850	4.61453	$5{,}89 \times 10^{-6}$
β_1	-0.029887	0.00448	-8.48864	$< 2 \times 10^{-16}$

Tabla 2: 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	$3{,}940854 \times 10^{-6}$
β_1	0.9706	0.9640	0.9773	$2{,}105584 \times 10^{-17}$

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:

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

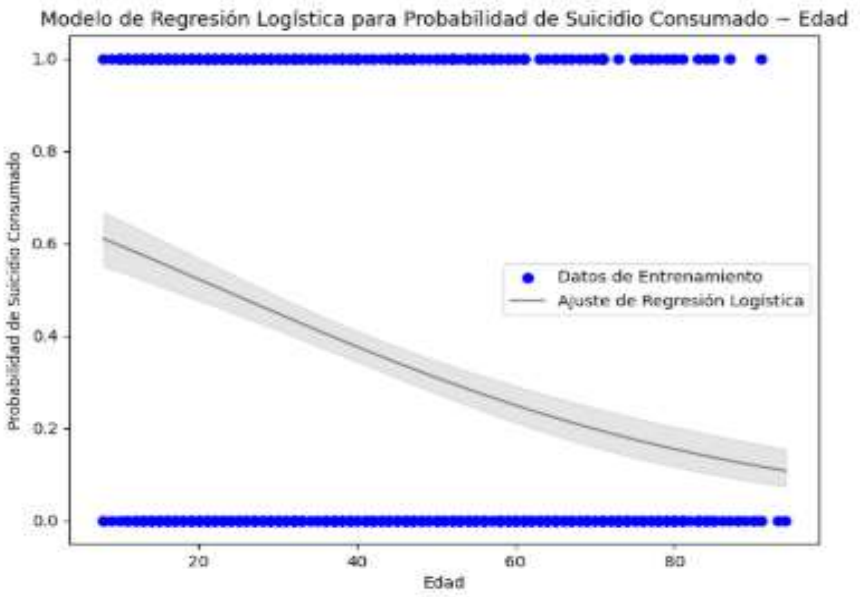


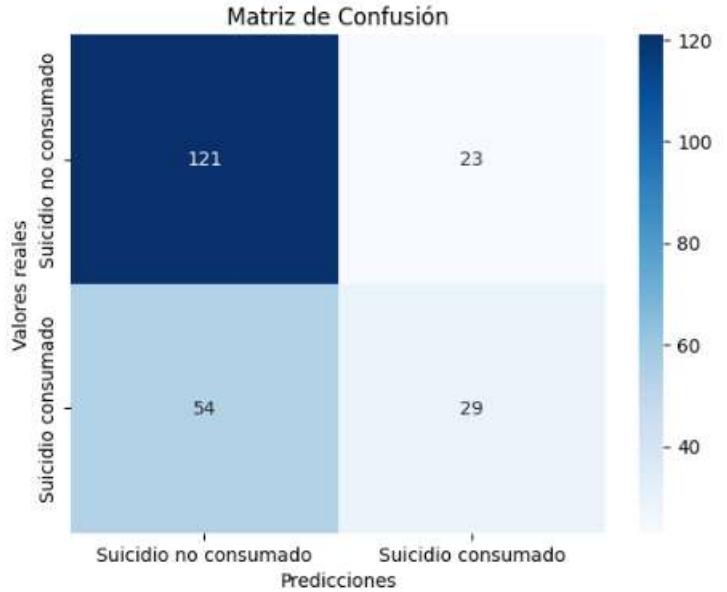
Gráfico del modelo con el intervalo de confianza

RESULTADOS

Evaluación del modelo

Matriz de confusión

La matriz de confusión es una herramienta fundamental en la evaluación de modelos de clasificación. Proporciona un resumen detallado del rendimiento del modelo al comparar sus predicciones con los valores reales en un conjunto de datos.



Matriz de confusión para el modelo de Regresión Logística Simple

El modelo es capaz de clasificar correctamente el 66 % de las observaciones cuando se emplean los datos de prueba.
Donde, este porcentaje de clasificación se denomina Accuracy y se obtiene de la siguiente manera:

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

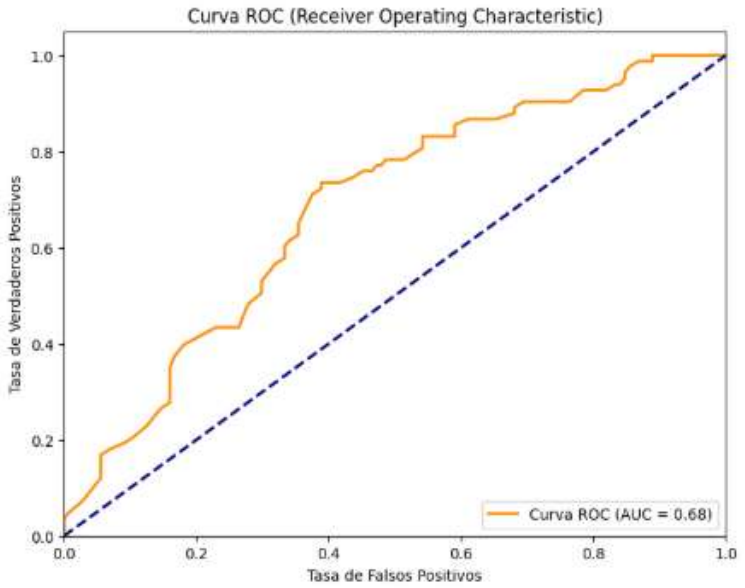
RESULTADOS

Evaluación mediante Curva ROC y AUC

La Curva ROC es una representación gráfica de la capacidad de discriminación de un modelo de clasificación binaria en diferentes umbrales de decisión.

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

El Área Bajo la Curva (AUC) es una métrica utilizada para evaluar la calidad de un clasificador binario a través de la curva ROC.



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.

Tabla 1: Coeficientes de regresión logística estimados

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	$6{,}53 \times 10^{-06}$
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 2: Intervalos de confianza para los Odds Ratio de β_i

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}

Los resultados resaltan la importancia de la detección temprana y la intervención preventiva en poblaciones con factores de riesgo identificados, como los antecedentes de intentos previos de suicidio y la remisión al servicio de psiquiatría.

RESULTADOS

En resumen tenemos:

- Los valores ajustados vienen dados por la ecuación:

$$\hat{\pi} = \frac{1}{e^{-(\widehat{\beta_0}+\widehat{\beta_1}\cdot Sexo_{Mujer}+\widehat{\beta_2}\cdot Edad+\widehat{\beta_3}\cdot Intentos\ Previos_{Si}+\widehat{\beta_4}\cdot Psiquiatría_{si})}}$$

- El Logit estimado viene dado por la ecuación:

$$\widehat{Logit}(\pi) = 1.69 - 0.74 \cdot Sexo_{Mujer} - 0.033 \cdot Edad - 0.71 \cdot Intentos\ Previos_{Si} - 1.62 \cdot Psiquiatría_{si}$$

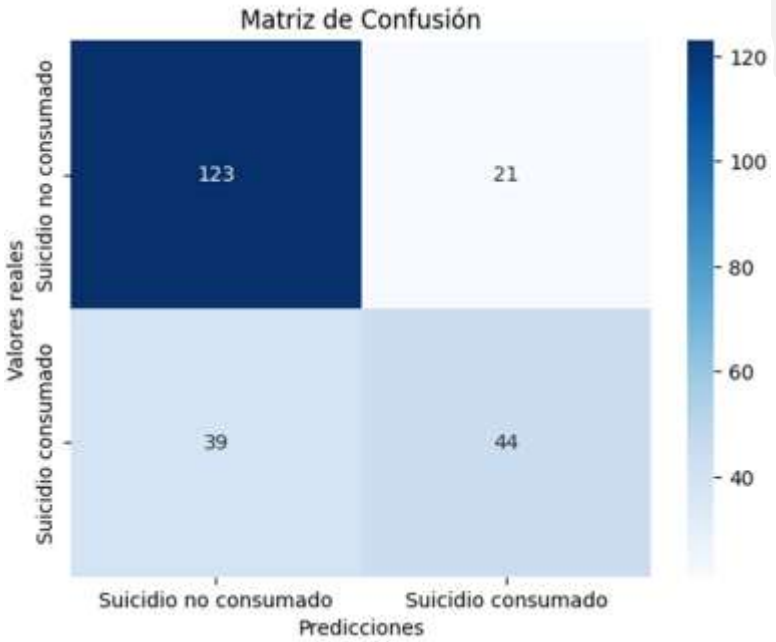
RESULTADOS

Evaluación del modelo

Matriz de confusión

Los resultados que se obtienen la figura se basan en los datos destinados para evaluar el modelo, los cuales constituyen un 20% del total.

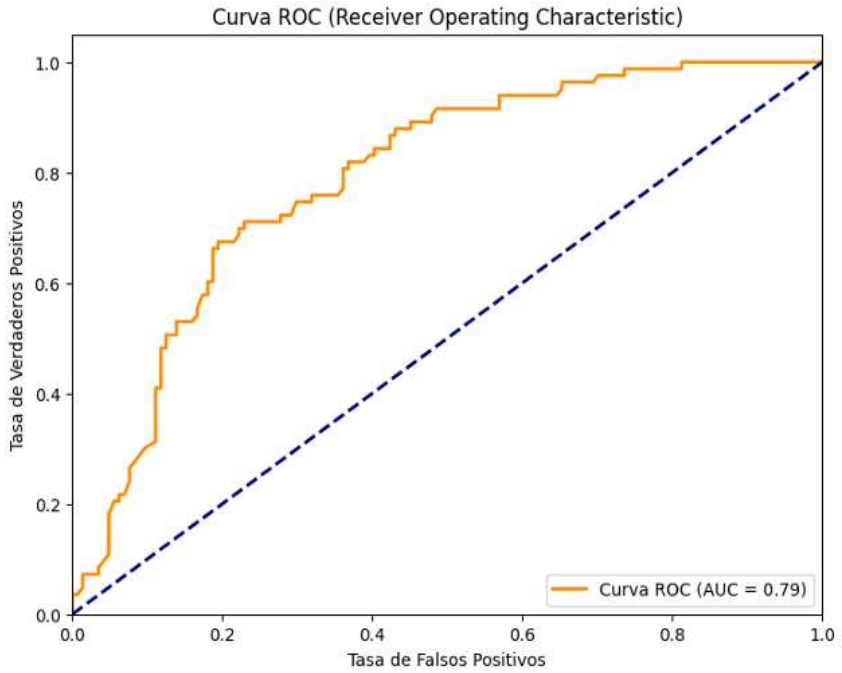
El modelo es capaz de clasificar correctamente el 74% de las observaciones cuando se emplean los datos de prueba. Recuerde que este porcentaje de clasificación se denomina Accuracy



RESULTADOS

Evaluación mediante Curva ROC y AUC

El modelo de regresión logística múltiple, 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%.



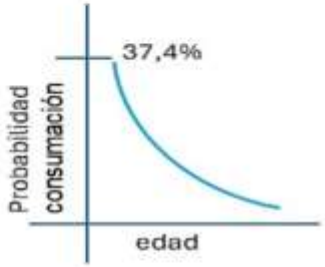
DISCUSIÓN DE RESULTADOS

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 la Edad, proporciona una mejor comprensión de los factores asociados con el riesgo de suicidio consumado.

ENCUENTRO DEPARTAMENTAL DE SEMILLEROS DE INVESTIGACIÓN

CONCLUSIONES

- **El presente estudio es de gran relevancia, conjuga conceptos matemáticos, probabilísticos y de estadística inferencial asociados a la regresión logística para el estudio del suicidio, problema de interés social.**

- **Se subraya la importancia de continuar investigando y desarrollando estrategias de prevención del suicidio, enfocadas en las poblaciones más vulnerables.**

- **Para otros trabajos, es de gran valor complementar con técnicas avanzadas de aprendizaje automático:**
- **- Mejorar predicciones**
- **- Identificar patrones.**
- **- Mayor precisión.**
- **- Mejorar las intervenciones preventivas.**

IMPACTOS

A través de este trabajo se espera poder contribuir con la investigación asociada al suicidio, problema que aqueja a nuestra sociedad, necesitando el apoyo de los gobiernos para realizar campañas en la prevención y tratamiento.

Esta teoría y aplicación es relevante y con seguridad será la base para estudios posteriores utilizando herramientas de machine learning para mejorar el modelo y tener resultados más robustos.

Referencias Bibliográficas

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

[2] [Castañeda, 2004] Castañeda, L. B. (2004). Probabilidad. Univ. Nacional de Colombia.

[3] [Chowdhury, 2021] Chowdhury, S. M. R. (2021). A review of logistic regression and its application.

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

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

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

Gracias

XVIII ENCUENTRO DEPARTAMENTAL DE SEMILLEROS DE INVESTIGACIÓN