Riesgo de Hipertensión

“Análisis y Predicción del Riesgo de Hipertensión”

Autor: Torres Jorges, David Jesus

Fecha: 14/08/2025

Repositorio: DavidJesusTJ - Overview

Resumen Ejecutivo

Objetivo del estudio

El presente proyecto tiene como finalidad, desarrollar un análisis de datos completo, iniciando con la exploración hasta el modelado de los datos, haciendo uso de un conjunto de datos simulado pero a la vez realista dada las variables que se relación con el problema real, los cuales afectan a la predicción de la hipertensión. Se busca identificar factores clave de riesgo mediante análisis estadístico y construir modelos de clasificación robustos, los cuales permitan evaluar la probabilidad de hipertensión en función de indicadores demográficos, médicos y de estilo de vida.

Puntos clave del EDA

- Impacto directo de las variables como “Fumador”, “Estado anterior con hipertensión” y “Familiares con hipertensión”, generan una relación fuerte con el estado de hipertensión actual del paciente.
- Por otro lado las personas de edad avanzada son las que en mayoría tienen hipertensión al igual que las personas que tienen un alto nivel de estrés.

Principales hallazgos y rendimiento de modelos

- Se implementó el modelo de Regresión Logística, dado que en conjuntos de datos médicos es más valioso tener una probabilidad más cercano a lo real, por ende se evaluó los supuestos que requiere este modelo para su correcto funcionamiento, cumpliendo con cada uno de ellos (independencia, multicolinealidad, linealidad, outliers, tamaño del dataset).
- Posteriormente se evaluó el modelo con los más adecuados parámetros, dando resultados prometedores, como una tasa de correcta predicción en el 82% de los casos, así como un 63% por encima de una predicción aleatoria.

Conclusión general y relevancia del análisis

Este análisis lo que muestra es que aún usando un conjunto de datos simulado pero muy bien diseñado, es posible construir un modelo de clasificación robusto para predecir el riesgo de hipertensión. Lo que los datos muestran es que se pueden identificar variables clave que generan impacto en la predicción. Este estudio aunque es trabajado con datos no reales tiene relevancia académica, pues permite mostrar el flujo de trabajo de ciencia de datos desde el análisis exploratorio hasta el modelado predictivo en el contexto clínico o de salud pública.

Introducción

Contexto del problema

La hipertensión arterial es uno de los principales factores de riesgo para la aparición de enfermedades cardiovasculares, incluyendo accidentes cerebrovasculares y enfermedad renal. Además, es altamente prevalente y puede progresar sin síntomas visibles, lo cual la convierte en un problema de salud pública prioritario a nivel global Lippincott.

En este proyecto, se aplica exclusivamente un modelo de Regresión Logística, técnica ampliamente utilizada para la predicción en salud, permitiendo estimar la probabilidad de hipertensión a partir de variables clínicas, demográficas y de estilo de vida. Esta metodología ha sido respaldada por estudios que demuestran su utilidad como modelo base en contextos de predicción cardiovascular arXiv.

Objetivo general:

Evaluar el rendimiento predictivo de un modelo de regresión logística para identificar la presencia de hipertensión en función de las variables en el dataset.

Objetivos específicos:

- Realizar un análisis exploratorio de datos (EDA) para entender la distribución y relación de variables.
- Entrenar y evaluar un modelo de regresión logística y reportar sus métricas clave.
- Interpretar los coeficientes del modelo para identificar factores relevantes asociados a la hipertensión.
- Documentar limitaciones y puntos de mejora del análisis para futuros estudios.

Preguntas clave del análisis:

- ¿Qué variables muestran mayor asociación estadística con la hipertensión?
- ¿Con qué precisión la regresión logística predice la presencia de hipertensión en este dataset?
- ¿Qué aportan los coeficientes del modelo en la interpretación clínica del riesgo?
- ¿Cuáles son las limitaciones específicas de este enfoque aplicado al dataset?

Descripción de los Datos

Fuente del dataset

El dataset utilizado proviene de Kaggle y se denomina *"Hypertension Risk Prediction Dataset"*. Contiene 1,985 registros con 11 variables significativas, construidas a partir de conocimientos clínicos y patrones de epidemiología en salud pública kaggle.com.

Tipos de variables

- **Numéricas:** Continuas o discretas.
- **Categóricas:** Variables con categorías ordinales o nominales.
- **Binarias:** Indicadores dicotómicos.

Diccionario de Variables

A continuación, se presenta una tabla con los nombres, tipos y descripciones de las variables en el dataset:

Variable	Tipo	Descripción
Age	Discreta	Edad del individuo en años
Salt_Intake	Continua	Consumo diario de sal
Stress_Score	Discreta	Nivel de estrés
BP_History	Ordinal	Estado anterior de la presión
Sleep_Duration	Continua	Promedio de horas de sueño
BMI	Continua	Índice de masa corporal
Medication	Nominal	Tipo de medicación
Family_History	Binaria	Historia familiar de hipertensión
Exercise_Level	Ordinal	Nivel de actividad física
Smoking_Status	Binaria	Si el paciente es fumador o no
Has_Hypertension	Binaria	Presencia de hipertensión (Sí/No)

Análisis Exploratorio de Datos

En esta etapa se examinó la estructura y el comportamiento de las variables del dataset con el objetivo de identificar patrones, relaciones y posibles problemas en los datos. Mencionar que solo la variables de *Medication* tiene valores nulos que fueron reemplazados.

Distribución de variables:

Prueba para evaluar si las variables tienen distribución normal.

```
Variable: Age | Estadístico: 0.9528587316597226 | P-valor: 8.476822953691877e-25 | Se rechaza la H0
Variable: Salt_Intake | Estadístico: 0.9991122093878423 | P-valor: 0.4547340639439232 | No se rechaza la H0
Variable: Stress_Score | Estadístico: 0.9390180675264523 | P-valor: 8.022837677574015e-28 | Se rechaza la H0
Variable: Sleep_Duration | Estadístico: 0.9989148243096209 | P-valor: 0.26622888090860736 | No se rechaza la H0
Variable: BMI | Estadístico: 0.9994453066490963 | P-valor: 0.8604507365925207 | No se rechaza la H0
```

*En estos datos sólo presentan normalidad **Salt_Intake**, **Sleep_Duration** y **BMI**, se podría evaluar usar una transformación de la z-score.*

Prueba para evaluar igualdad de clases en variables categóricas.

```
BP_History: Estadístico Chi² = 42.8746 | p-valor = 0.0000 → Se rechaza H₀: proporciones distintas
Medication: Estadístico Chi² = 83.9292 | p-valor = 0.0000 → Se rechaza H₀: proporciones distintas
Family_History: Estadístico Chi² = 0.1134 | p-valor = 0.7364 → No se rechaza H₀: proporciones iguales
Exercise_Level: Estadístico Chi² = 224.4826 | p-valor = 0.0000 → Se rechaza H₀: proporciones distintas
Smoking_Status: Estadístico Chi² = 363.1239 | p-valor = 0.0000 → Se rechaza H₀: proporciones distintas
Has_Hypertension: Estadístico Chi² = 3.1441 | p-valor = 0.0762 → No se rechaza H₀: proporciones iguales
```

*Las pruebas muestran que las proporciones de las clases en las variables son significativamente diferentes en el caso de **BP_History**, **Medication**, **Exercise_Level** y **Smoking_Status.***

Análisis univariante:

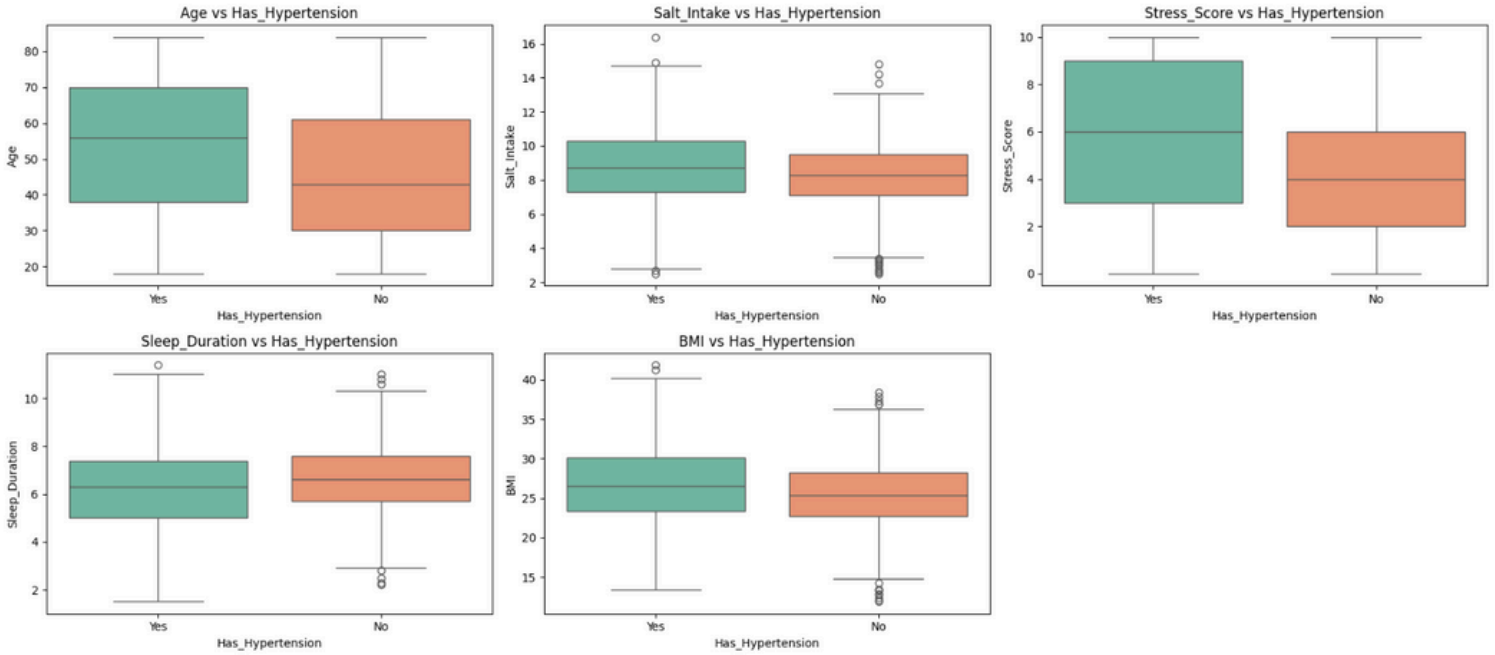
La variable más dispersa es la de *Strees_Score* y la que menos dispersión tiene es la de *Sleep_Duration*, viendo estos datos será necesario una estandarización de los datos.

	Age	Salt_Intake	Stress_Score	Sleep_Duration	BMI
count	1985.000000	1985.000000	1985.000000	1985.000000	1985.000000
mean	50.341058	8.531688	4.979345	6.452242	26.015315
std	19.442042	1.994907	3.142303	1.542207	4.512857
min	18.000000	2.500000	0.000000	1.500000	11.900000
25%	34.000000	7.200000	2.000000	5.400000	23.000000
50%	50.000000	8.500000	5.000000	6.500000	25.900000
75%	67.000000	9.900000	8.000000	7.500000	29.100000
max	84.000000	16.400000	10.000000	11.400000	41.900000

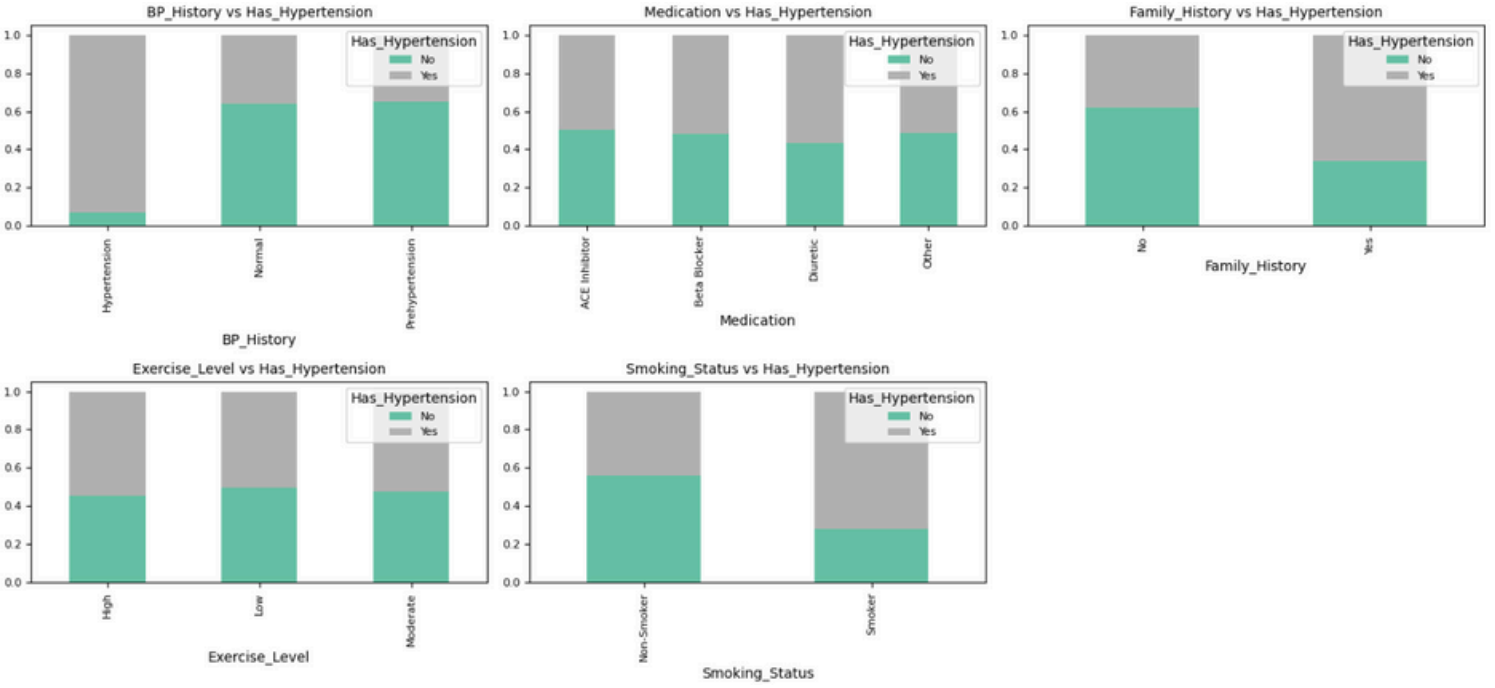
En los datos se tiene que la mayoría de ellos tienen un historial anterior de la presión arterial en *Normal*, la medicación más usada es *Beta Blocker*, el historial familiar de tener hipertensión es *No*, el nivel de ejercicio es de *Low*, el estados de fumador es de *Non-Smoker* y la mayor parte de ellos tiene hipertensión.

	BP_History	Medication	Family_History	Exercise_Level	Smoking_Status	Has_Hypertension
count	1985	1186	1985	1985	1985	1985
unique	3	4	2	3	2	2
top	Normal	Beta Blocker	No	Low	Non-Smoker	Yes
freq	796	412	1000	936	1417	1032

Análisis bivariante:

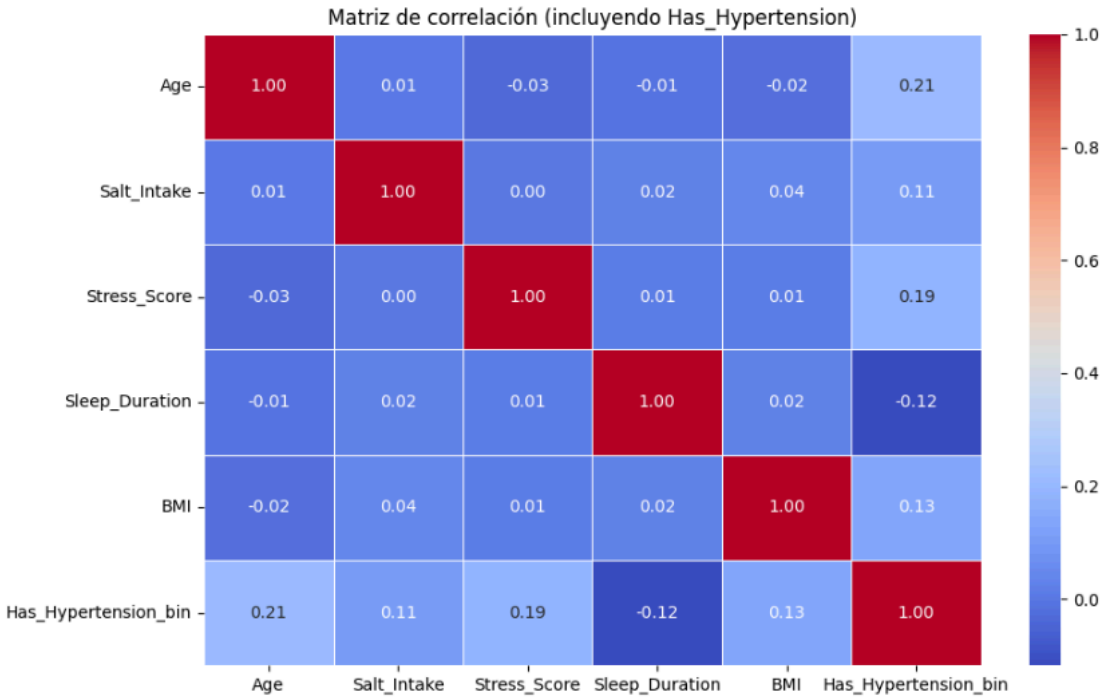


Aquí las variables que más representan una diferencia entre una persona que tiene hipertensión o no son la de **Age** y **Stress_Score**.

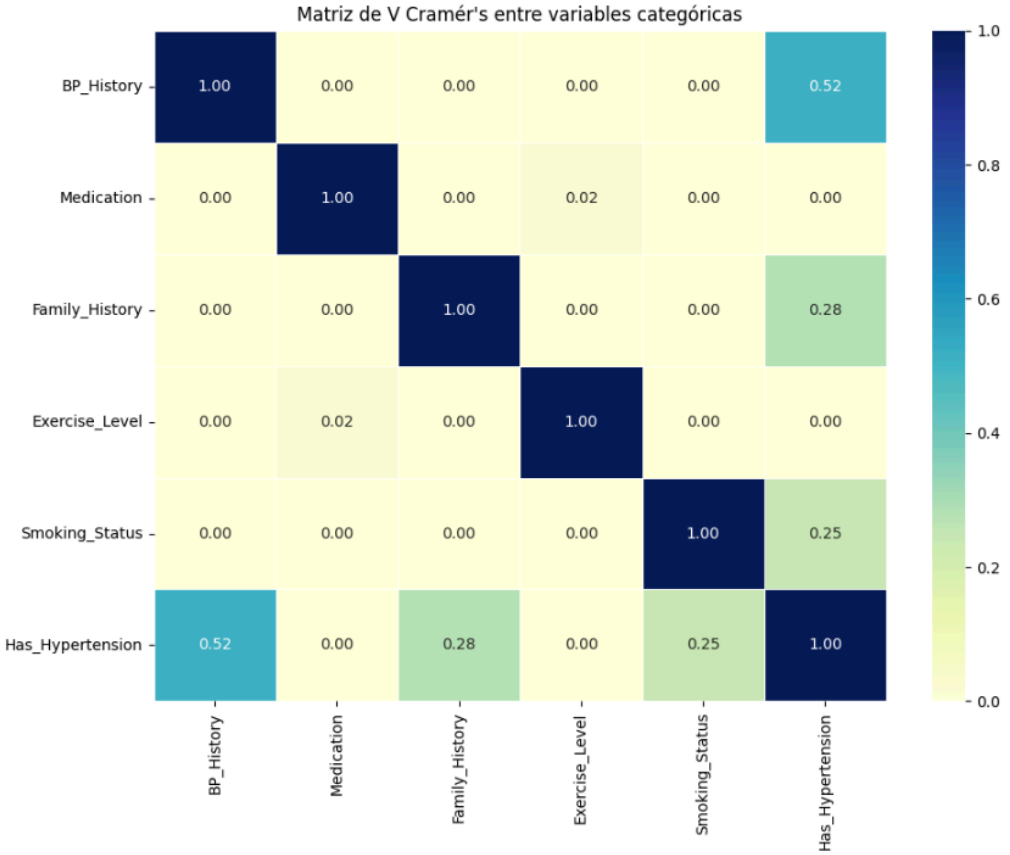


En estos patrones se aprecia que el paciente que en el historial anterior ha tenido hipertensión es muy probable que lo tenga ahora, por otro lado si tiene familiares con hipertensión también es probable que el lo tenga, adicional a ello, si el paciente es fumador también es más probable que tenga hipertensión.

Correlaciones:



Entre las variables numéricas no se presenta multicolinealidad, aunque las variables tienen correlaciones bajas con el target, esta relación podría incrementarse con la interacción de variables.



Entre las variables categóricas no hay multicolinealidad tampoco, sin embargo, se presentan que hay variables que tienen relación fuerte con el target, de igual manera en las interacciones con las demás variables estas relaciones podrían variar.

Preprocesamiento de Datos

Para garantizar la calidad y coherencia de la información antes del modelado, se llevaron a cabo las siguientes transformaciones y ajustes en los datos:

Cambio de tipos de datos:

Se ajustaron los tipos de datos para asegurar su correcto tratamiento en las etapas posteriores de análisis y modelado. En **Age**, **Salt_Intake**, **Stress_Score**, **Sleep_Duration** y **BMI** como *float64.* Para **BP_History**, **Medication**, **Family_History**, **Exercise_Level**, **Smoking_Status** y **Has_Hypertension** como *category*.

Escalado de variables numéricas:

Las variables **Age**, **Salt_Intake**, **Stress_Score**, **Sleep_Duration** y **BMI** fueron escaladas para homogeneizar sus rangos y mejorar la eficiencia del algoritmo de regresión logística. Al observar que se tiene unos valores atípicos para poder sobrellevarlos sin problemas se aplicó el escalado de datos por *RobustScaler*.

Codificación de variables categóricas:

- Para variables **ordinales** como *BP_History* y *Exercise_Level*, las categorías fueron reemplazadas por valores numéricos 0, 1 y 2, respetando el orden natural de sus niveles.
- Las variables **binarias** *Family_History*, *Smoking_Status* y *Has_Hypertension* se mapean a valores 0 y 1 según su condición (No/Yes, Non-Smoker/Smoker).
- Para la variable **nominal** *Medication*, se aplicó *one-hot encoding* para crear variables indicadoras.

Tratamiento de valores atípicos:

Se detectaron outliers mediante métodos gráficos, pero dado que su número era reducido, se decidió mantenerlos para no perder información relevante.

	Q1	Q3	IQR	Límite inferior	Límite superior	Cantidad de outliers	Porcentaje
Salt_Intake	-0.481481	0.518519	1.0	-1.981481	2.018519	17.0	0.86
BMI	-0.475410	0.524590	1.0	-1.975410	2.024590	16.0	0.81
Sleep_Duration	-0.523810	0.476190	1.0	-2.023810	1.976190	12.0	0.60
Age	-0.484848	0.515152	1.0	-1.984848	2.015152	0.0	0.00
Stress_Score	-0.500000	0.500000	1.0	-2.000000	2.000000	0.0	0.00

Separación en conjuntos de entrenamiento y prueba:

El conjunto de datos se dividió en 75% para entrenamiento y 25% para prueba, utilizando estratificación para preservar la proporción original de la variable objetivo ***Has_Hypertension.***

Modelado

En esta etapa se empleó **únicamente** el modelo de **Regresión Logística** para predecir la probabilidad de que un individuo presente hipertensión, dado que se trata de un problema de clasificación binaria y el objetivo es interpretar los efectos de cada predictor en la variable objetivo.

Verificación de supuestos del modelo

Antes del ajuste, se verificaron los principales supuestos de la regresión logística:

1. **Independencia de observaciones:** se confirmó que cada registro corresponde a un individuo distinto.
2. **Ausencia de multicolinealidad extrema:** se evaluó la correlación entre predictores, sin detectar valores críticos de VIF.
3. **Linealidad del logit:** se comprobó que las variables numéricas mantienen una relación lineal con el log-odds de la variable objetivo.
4. **Ausencia de outliers influyentes:** aunque se identificaron algunos valores atípicos, su número fue reducido y no se eliminan para evitar pérdida de información.
5. **Tamaño muestral suficiente:** razón calculada entre el número de predictores vs el número de registros.

Evaluación global del modelo

Medidas de Pseudo R^2:

- Cox & Snell: **0.4936**
- Nagelkerke: **0.6585**
- McFadden: **0.4914**
- Basado en devianza: **0.4914**

Estos valores indican que el modelo explica una proporción considerable de la variabilidad de la respuesta.

Prueba de verosimilitud (Likelihood Ratio Test):

- p-valor = $\mathbf{3.51 \times 10^{-208}}$

Se rechaza la hipótesis nula de que el modelo con predictores no mejora respecto al modelo nulo, concluyendo que el modelo es altamente significativo.

Desviancia:

- Modelo con predictores: **1047.83**
- Modelo nulo: **2060.39**

La reducción sustancial de la desviancia confirma el mejor ajuste del modelo completo.

Criterios de información:

- AIC (modelo completo): **1075.83** vs. nulo: **2062.39**
- BIC (modelo completo): **1150.10** vs. nulo: **2067.69**

Ambos criterios favorecen el modelo con predictores, evidenciando un equilibrio óptimo entre ajuste y complejidad.

Coeficientes, Odds Ratios e intervalos de confianza

Variable	Coef.	OR	IC 2.5%	IC 97.5%	p-valor
const ***	-3.83	0.0217	0.0122	0.0385	3.20×10^{-39}
Age ***	1.74	5.67	4.21	7.65	4.59×10^{-30}
Salt_Intake ***	0.78	2.19	1.77	2.72	8.46×10^{-13}
Stress_Score ***	1.87	6.46	4.62	9.03	9.15×10^{-28}
BP_History ***	2.29	9.85	7.63	12.71	6.00×10^{-69}
Sleep_Duration ***	-0.89	0.41	0.33	0.52	3.16×10^{-14}
BMI ***	0.83	2.30	1.85	2.87	1.05×10^{-13}
Family_History ***	2.31	10.07	7.06	14.35	2.42×10^{-37}
Exercise_Level	0.11	1.12	0.92	1.36	0.268
Smoking_Status ***	2.48	11.90	7.93	17.87	6.54×10^{-33}
Medication_Beta Blocker	-0.15	0.86	0.52	1.43	0.567
Medication_Diuretic	0.33	1.39	0.78	2.50	0.265
Medication_Not specified	0.00	1.00	0.64	1.58	0.991
Medication_Other	-0.12	0.88	0.48	1.63	0.693

Principales hallazgos

Factores de riesgo más relevantes:

- *Smoking_Status* (OR ≈ 11.90)
    - Las personas que fuman tienen **aproximadamente 11.9 veces más probabilidades** de desarrollar hipertensión que quienes no fuman, manteniendo constantes los demás factores.
- *Family_History* (OR ≈ 10.07)
    - Tener antecedentes familiares de hipertensión incrementa el riesgo **en más de 10 veces** comparado con quienes no tienen esa historia familiar.
- *BP_History* (OR ≈ 9.85)
    - Un historial previo de presión arterial elevada multiplica el riesgo de hipertensión por **casi 10 veces**.
- *Stress_Score* (OR ≈ 6.46)
    - Un puntaje alto de estrés se asocia con un riesgo **6.46 veces mayor** de hipertensión en comparación con niveles bajos de estrés.
- *Age* (OR ≈ 5.67)
    - A mayor edad, el riesgo de hipertensión aumenta, siendo **5.67 veces mayor** en los grupos de mayor edad frente a los más jóvenes.
- *BMI* (OR ≈ 2.30)
    - Un índice de masa corporal más alto (sobrepeso/obesidad) incrementa el riesgo **en 2.3 veces** respecto a personas con IMC saludable.
- *Salt_Intake* (OR ≈ 2.19)
    - Un consumo elevado de sal eleva el riesgo de hipertensión **en 2.19 veces** comparado con ingestas bajas.

Factor protector:

- *Sleep_Duration* (OR ≈ 0.41)
    - Dormir más horas actúa como factor protector, reduciendo el riesgo de hipertensión **en un 59%** (1 − 0.41), en comparación con quienes duermen menos.

Variables no significativas: *Exercise_Level*, tipos de medicación, no mostraron asociación estadística significativa en este conjunto de datos.

Resultados

Tras el entrenamiento inicial del modelo de **Regresión Logística** y la posterior optimización de hiperparámetros, se observaron mejoras consistentes en las métricas de rendimiento.

Ajuste de hiperparámetros antes y después

Métrica	Antes del tuning	Después del tuning
Precisión (0)	0.80	0.81
Recall (0)	0.81	0.82
F1-Score (0)	0.81	0.81
Precisión (1)	0.82	0.83
Recall (1)	0.81	0.82
F1-Score (1)	0.82	0.82
Accuracy	0.81	0.82
Macro Avg	0.81	0.82
Weighted Avg	0.81	0.82

Parámetros óptimos identificados:

- **Precisión de búsqueda máxima:** 0.6608
- **Hiperparámetros finales:** penalty_solver='l2__saga', max_iter=483, C=0.33998

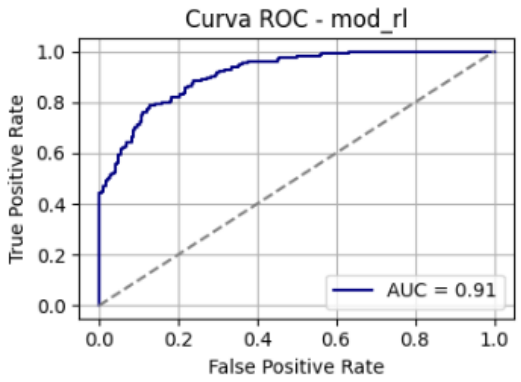
Evaluación de métricas adicionales

Métrica	Valor	Interpretación
Matthews CorrCoef	0.633	Buen equilibrio entre clases, considerando toda la matriz de confusión.
Cohen Kappa	0.633	Alto acuerdo con la clase real, corrigiendo el azar.
F2 Score	0.820	Alta sensibilidad, priorizando la detección de casos positivos.
Log Loss	0.370	Probabilidades moderadamente bien calibradas.
Brier Score	0.121	Excelente calibración de probabilidades predichas.
False Positive Rate	0.184	Moderado, aceptable en un contexto médico.
False Negative Rate	0.182	Muy bajo, más del 80% de casos positivos detectados correctamente.
Precision/Recall Ratio	1.012	Balance entre precisión y recall, con ligera inclinación hacia recall.

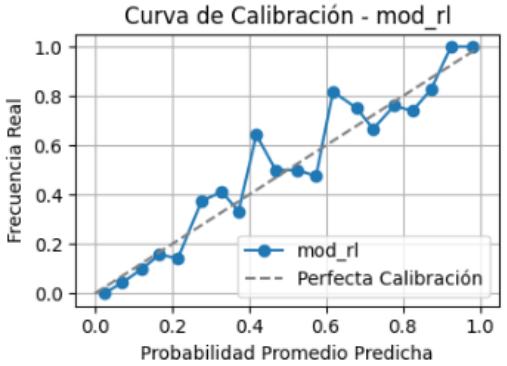
Visualizaciones del modelo

Se generaron las siguientes representaciones gráficas para evaluar el rendimiento:

Curva ROC:



Curva de calibración:



Curva Precision-Recall:

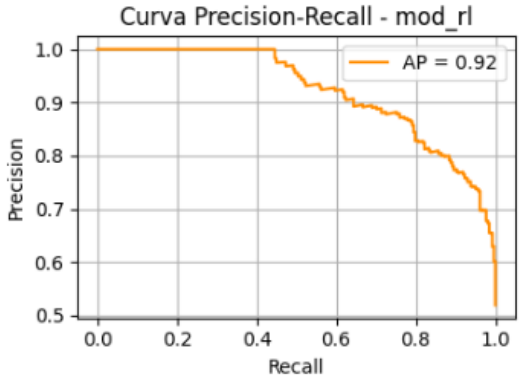
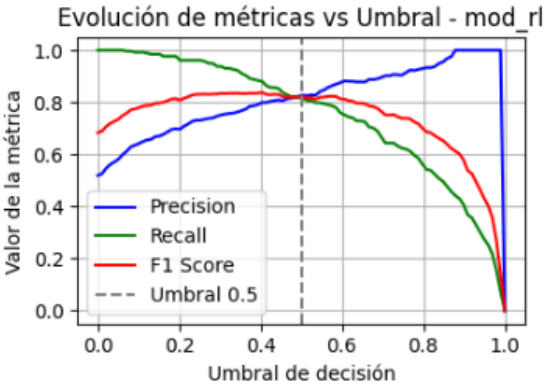
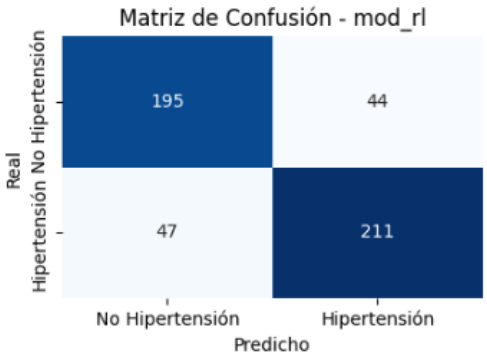


Gráfico de métricas vs. umbral de decisión:



Matriz de confusión:



Conclusión

Después de evaluar el modelo mod_rl sobre la data de test, considero que su desempeño es bastante sólido para el objetivo que perseguimos: identificar correctamente a personas con hipertensión.

Lo que más destaco es el F2 Score de 0.82, que refleja que el modelo está priorizando correctamente la detección de casos positivos, incluso si eso implica aceptar algunos falsos positivos. Esto es coherente con la naturaleza del problema, ya que en contextos de salud pública prefiero evitar falsos negativos, es decir, no pasar por alto a personas que realmente tienen hipertensión.

Además, el Matthews Correlation Coefficient (0.633) y el Cohen's Kappa (0.633) me indican que el modelo tiene un rendimiento balanceado y significativamente mejor que el azar, considerando todos los aspectos de la matriz de confusión, incluso si existiese cierto desbalance entre clases.

Por otro lado, la tasa de falsos negativos es baja (18.2%), lo cual considero un resultado muy positivo en este contexto. A pesar de que el modelo comete algunos falsos positivos (FPR ≈ 18.4%), es un compromiso aceptable si eso nos permite cubrir una mayor cantidad de pacientes que realmente están en riesgo.

Finalmente, tanto el log loss (0.370) como el Brier score (0.121) me indican que las probabilidades predichas son razonablemente bien calibradas, lo cual es útil si queremos usar estas salidas como insumo para decisiones clínicas o posteriores etapas de intervención.

En resumen, me siento conforme con el comportamiento del modelo sobre los datos de test, ya que cumple con el objetivo principal: maximizar la identificación de personas con hipertensión sin comprometer seriamente la precisión ni el balance general del modelo.