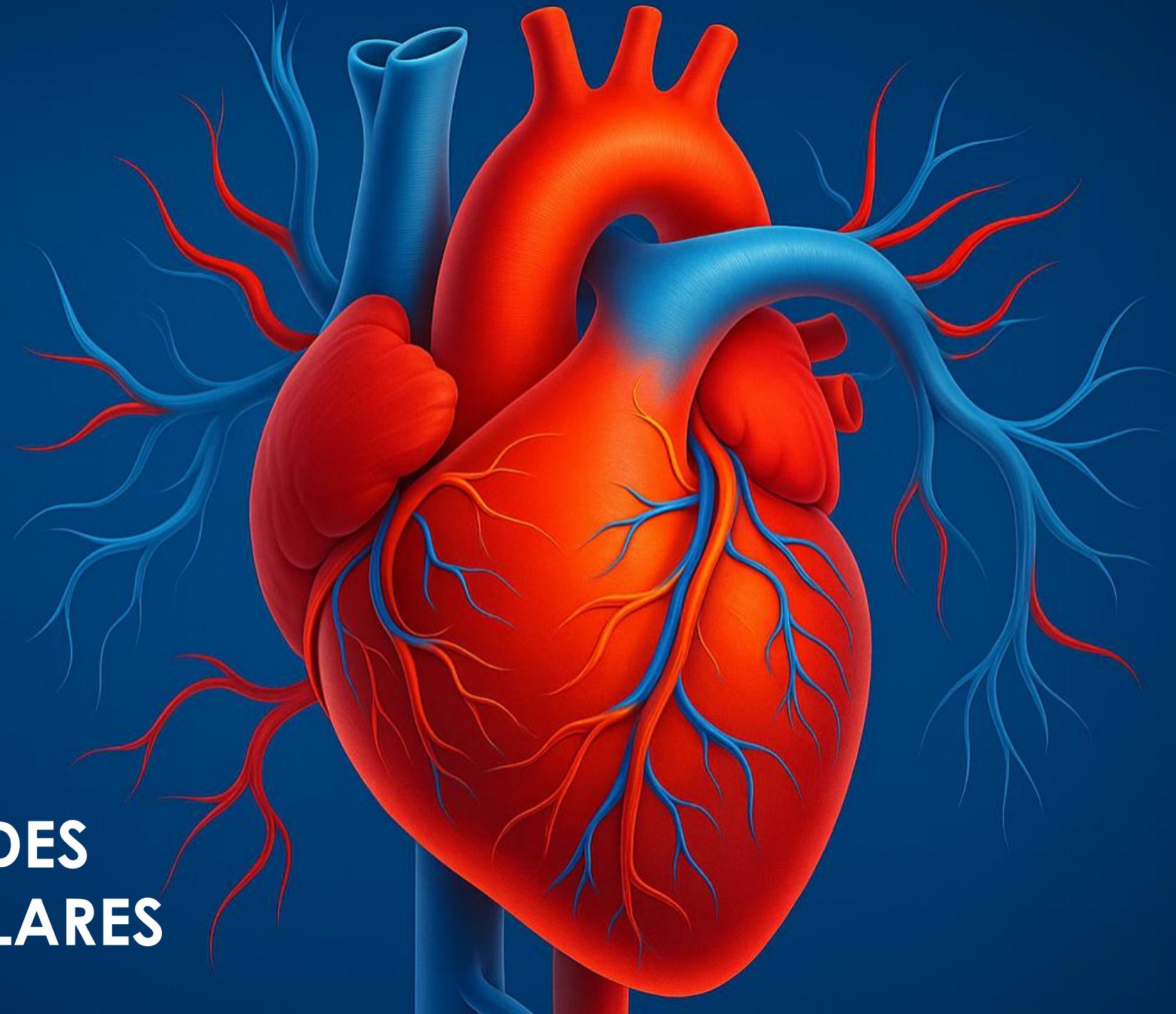


GRUPO 5

NUCLIO DIGITAL SCHOOL

Máster en Data Science & AI

MODELO PARA LA DETECCIÓN DE ENFERMEDADES CARDIOVASCULARES



ÍNDICE DEL TRABAJO

-  1. STORYTELLING
-  2. PREVALENCIA DE ENFERMEDADES
-  3. CONTEXTO DEL ESTUDIO
-  4. OBJETIVOS DEL ESTUDIO
-  5. OBTENCIÓN DE LOS DATOS
-  6. FLUJO DE TRABAJO
-  7. ETAPA INICIAL DEL ESTUDIO
-  8. EDA + PREPROCESAMIENTO
-  9. MODELOS DE CLASIFICACIÓN
-  10. ANÁLISIS DISCRIMINANTE LINEAL (LDA)
-  11. XGBOOST
-  12. SOFTWARE CLÍNICO
-  13. CONCLUSIONES DEL ESTUDIO

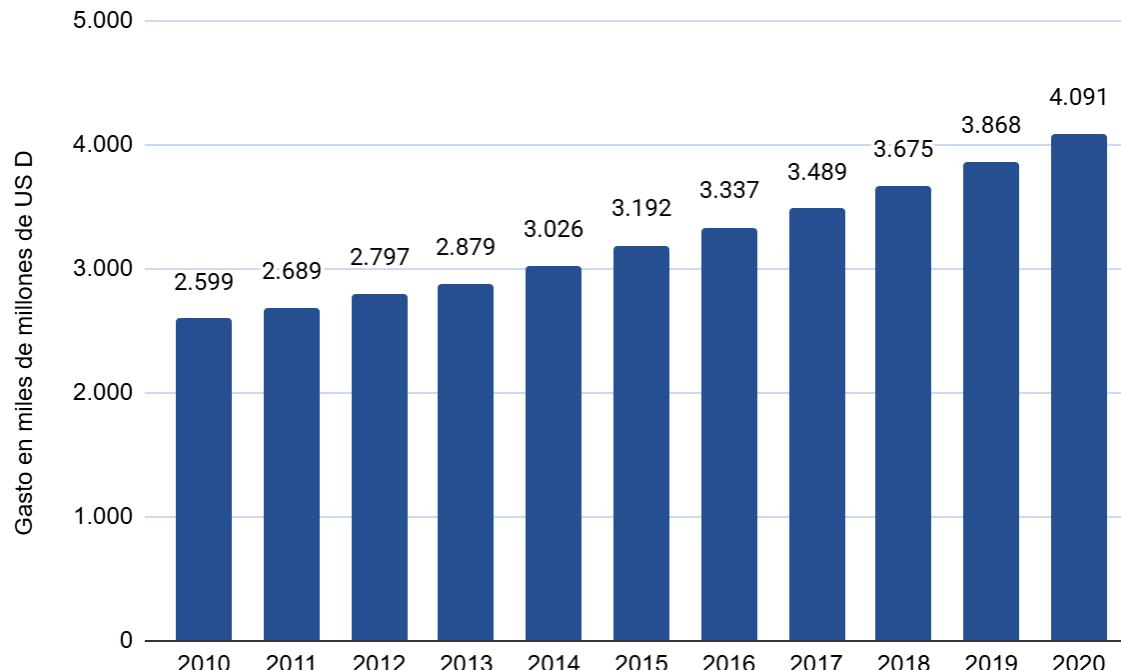


STORYTELLING

Gastos Médicos y Tipo de Financiación



- Evolución anual del gasto sanitario en EE.UU de 2010-2020



Fuente: Elaboración propia a partir de los datos de Statista

Las enfermedades cardíacas costaron alrededor de \$252.2mil
millones entre 2019 y 2020

- Principales fuentes de financiación en el GASTO

TOTAL:

- Programas públicos



- Pago directo (gasto de bolsillo)



- Sector privado

Líderes del mercado de seguros médicos y de salud de EE. UU.

1 UnitedHealth Group

2 Anthem

3 Humana Group

4 HCSC Group

5 Centene Corporation

STORYTELLING

Causas del Gasto Público y Beneficios del Modelo

- Causas del incremento del gasto sanitario:

Estructura **compleja** con aseguradoras y procedimientos de facturación =
GASTOS ADMINISTRATIVOS SIGNIFICATIVOS

Falta de acceso a la **prevención y atención primaria**

No hay criterios **homogéneos** de evaluación

Uso **inadecuado y excesivo** de servicios médicos



- Beneficios de la implementación del modelo para triaje aplicado en software clínico:

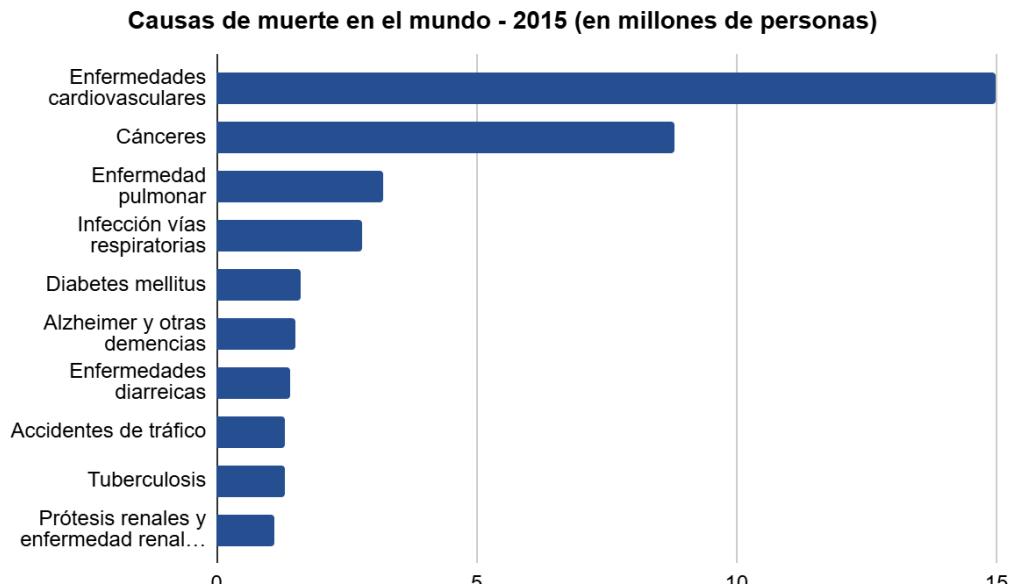
- **Reducción** de pruebas costosas.
- Menor “ruido” diagnóstico.
- Menos **pruebas** → menor **carga** en el sistema → más **agilidad** para casos críticos.
- Reducción de **costos asociados al tratamiento hospitalario**, debido a la optimización de los recursos.



Organización
Mundial de la Salud

PREVALENCIA DE ENFERMEDADES

Global (2015) y en U.S.A (2015-2020)



Fuente: Elaboración propia a partir de los datos la OMS (2024)

Number of Deaths for Leading Causes, US 2015-2020						
	2015	2016	2017	2018	2019	2020
Heart disease	633842	635260	647457	655381	659041	690882
Cancer	595930	598038	599108	599274	599601	598932
COVID-19						345323
Unintentional injuries	146571	161374	169936	167127	173040	192176
Stroke	140323	142142	146383	147810	150005	159050
Chronic lower respiratory diseases	155041	154596	160201	158486	156979	151637
Alzheimer disease	110561	116103	121404	122019	124614	133382
Diabetes	79535	80058	83564	84946	87647	101106
Influenza and pneumonia	57062	51537	55672	59120	49783	53495
Kidney disease	49959	50046	50633	51386	51565	52260
Suicide	44193	44965	47173	48344	47511	44834

Fuente: Elaboración propia a partir de los datos la CDC (2024)



Organización
Mundial de la Salud

CONTEXTO DEL ESTUDIO



American
Heart
Association®

CONTEXTO DEL ESTUDIO

AÑO 2019

Fuente: Organización Mundial
de la Salud. (2023)



68 %

Personas sanas dentro del rango
de bajo riesgo

VS

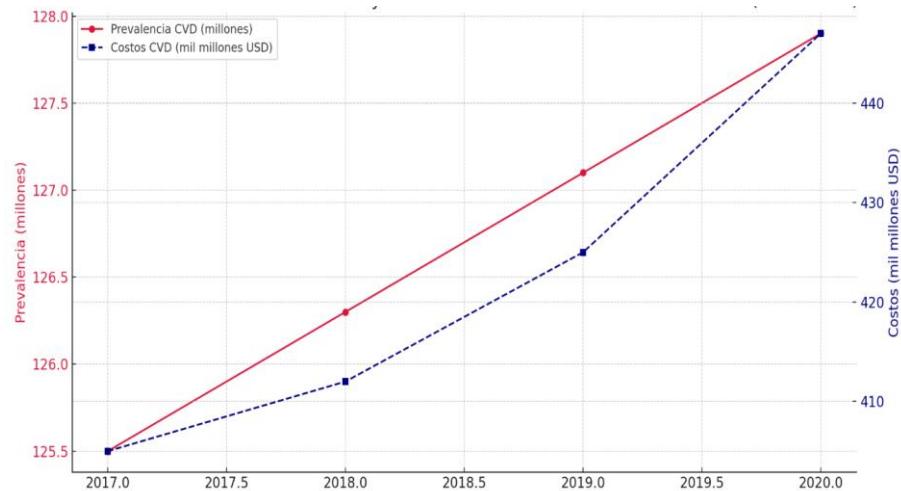
32 %

Personas con un alto riesgo o
condición cardiovascular que
requiere atención médica.

En los Estados Unidos:

- Las enfermedades cardíacas son la **principal causa de muerte** en hombres, mujeres y personas de la mayoría de los grupos raciales y étnicos.
- Una persona muere **cada 33 segundos** por enfermedad cardiovascular.

Tendencia estimada de Prevalencia y Costos de Enfermedades Cardiovasculares (2017-2020)



Fuente: Elaboración propia a partir de los datos de American Heart Association

OBJETIVOS DEL PROYECTO

Detección Precoz de Enfermedades Cardiovasculares

Optimizar la derivación a especialistas desde primaria

Recomendación de hábitos saludables

Seguimiento y monitoreo de variables (smartwatch)

Petición de pruebas de forma más eficiente

Gestión histórica clínica en centros de salud

Actualización del modelo y los registros con más datos



OBTENCIÓN DE LOS DATOS

National Health and Nutrition Examination Survey 2015-2020



CUESTIONARIOS

Información de hábitos y conductas, enfermedades previas y medicación

EXAMENES MÉDICOS

Recopila evaluaciones físicas y pruebas médicas realizadas a los pacientes



LABORATORIO

Lista de variables de laboratorio analizadas en muestras biológicas (sangre y orina)

DEMOGRAFICOS

Variables demográficas y pesos muestrales de las tendencias en salud y nutrición en EE.UU.

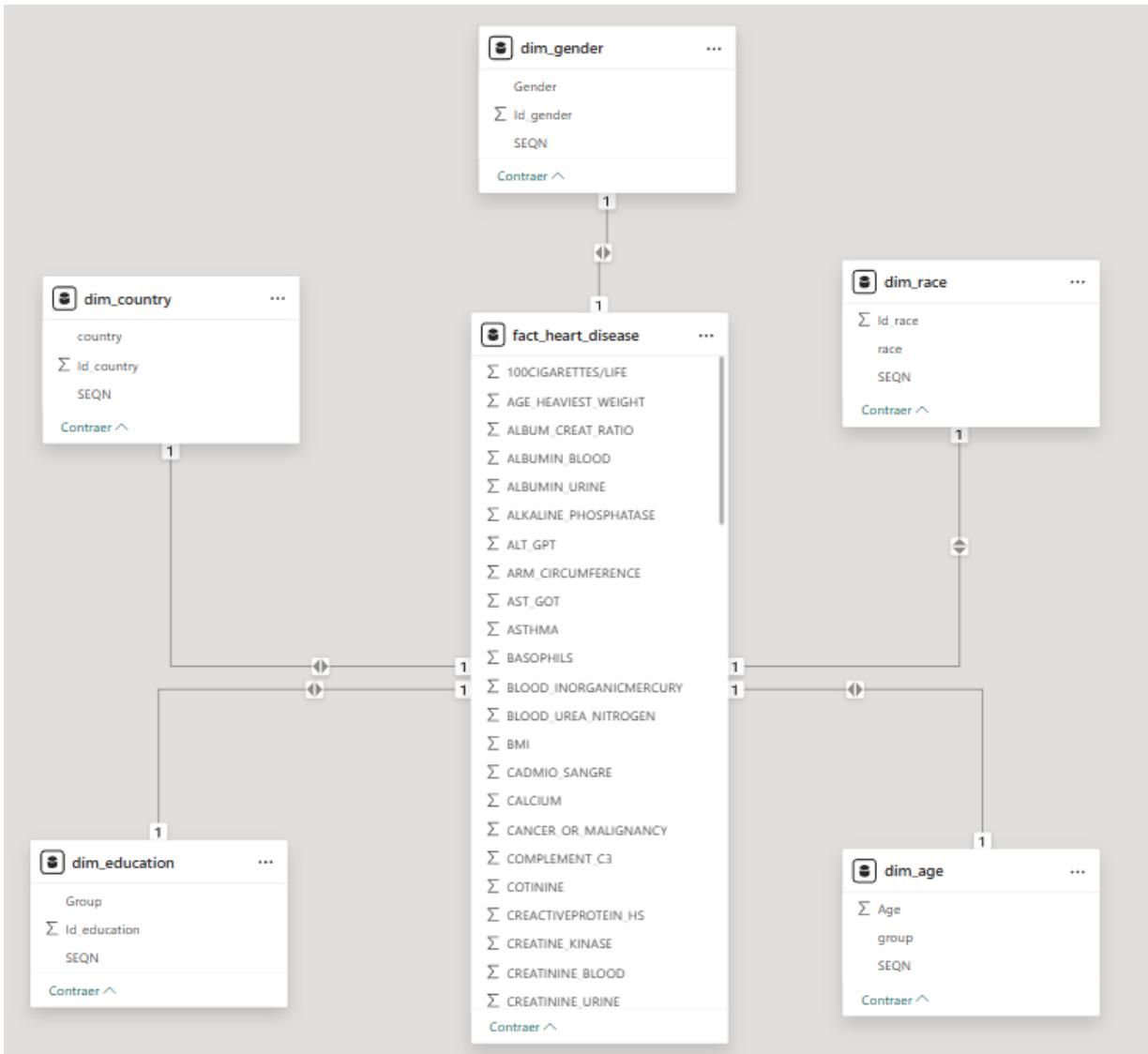


OBTENCIÓN DE LOS DATOS

National Health and Nutrition Examination Survey 2015-2020

OBTENCIÓN DE LOS DATOS

MODELO DE DATOS



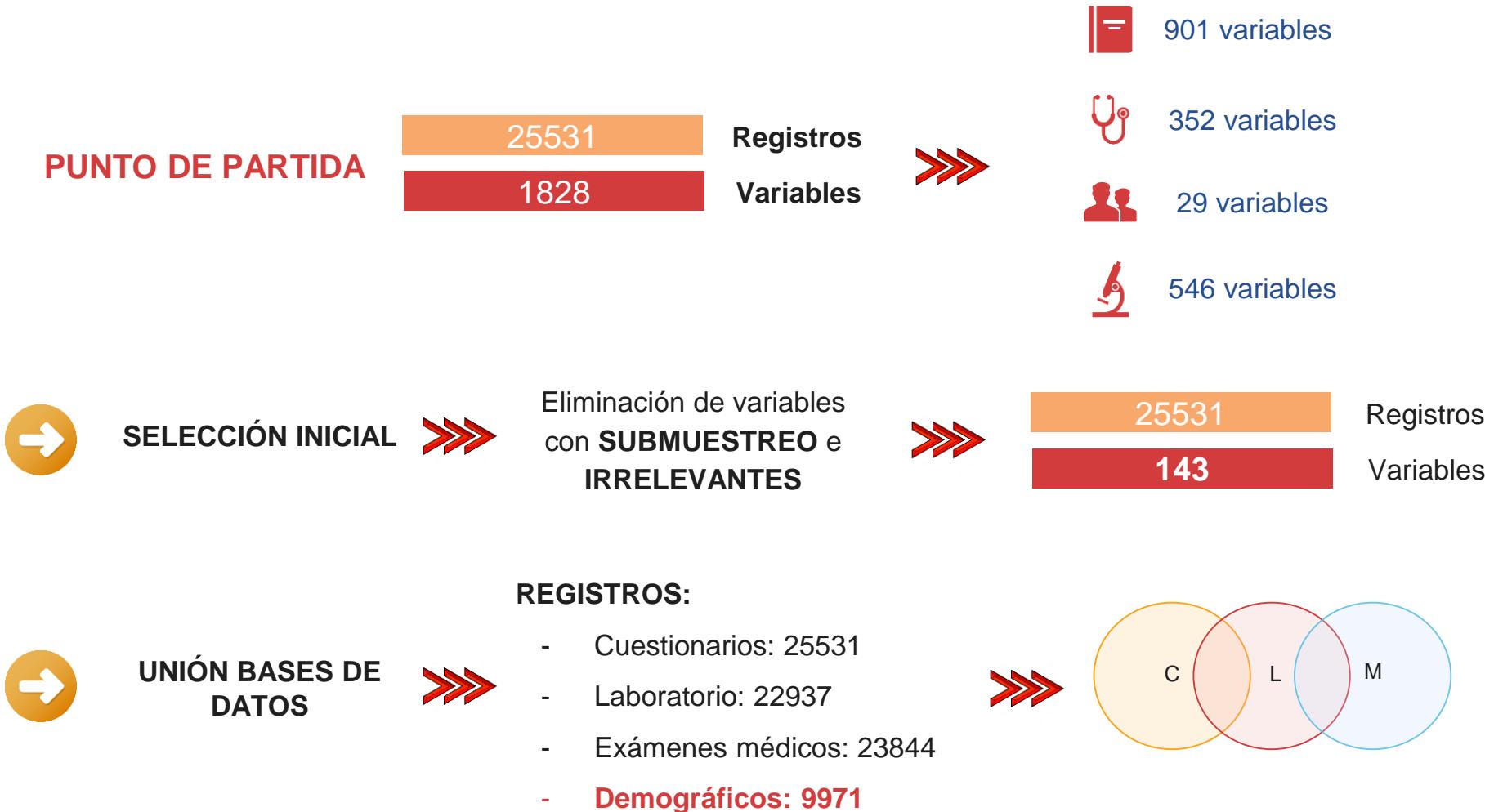
FLUJO DE TRABAJO

Etapas para el Desarrollo del Software



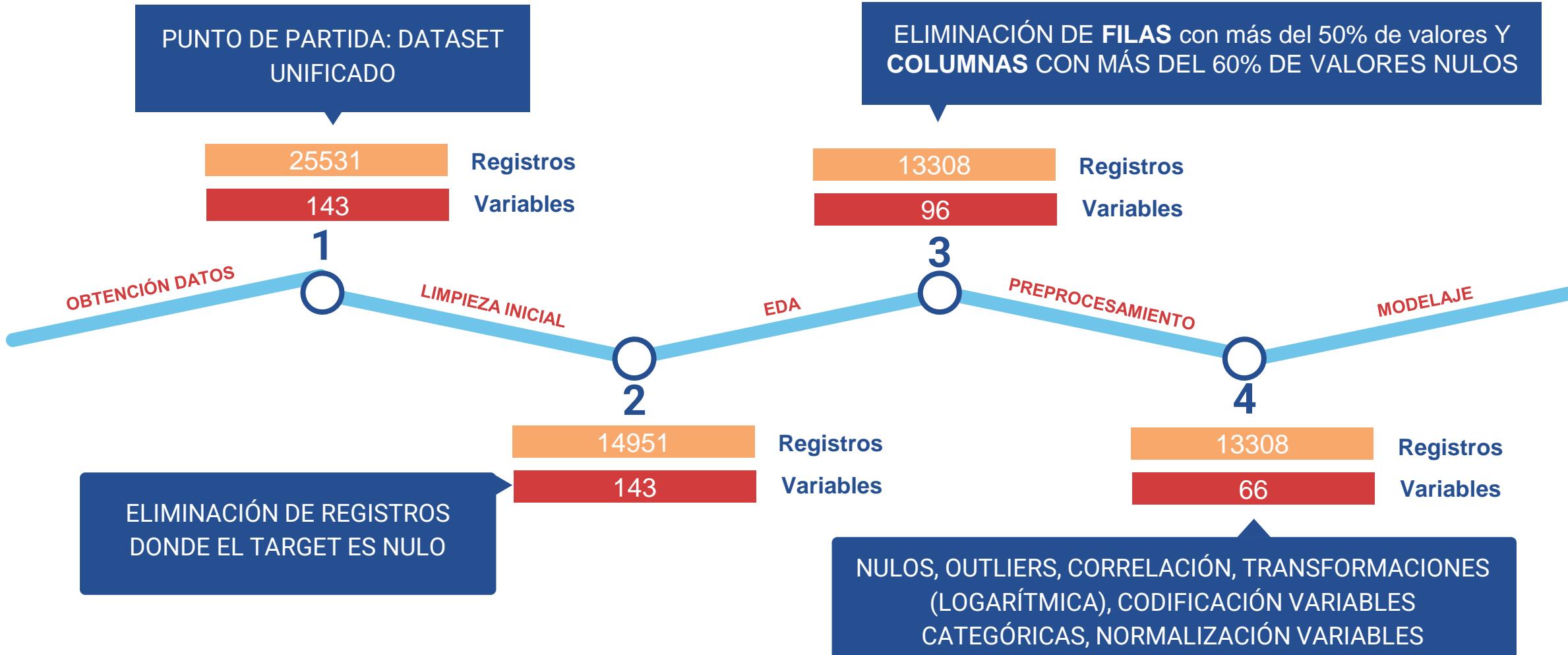
ETAPA INICIAL DEL ESTUDIO

Búsqueda y Recolección de Datos. Creación del Dataframe



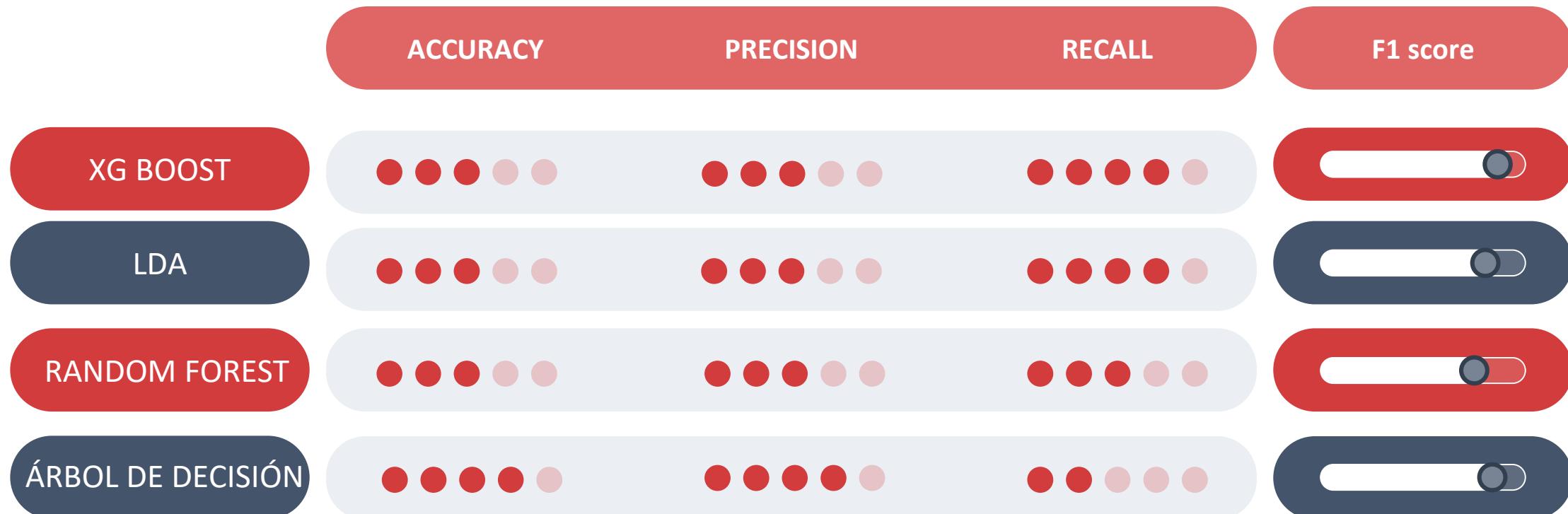
EDA + PREPROCESAMIENTO

Obtención del Dataset para el Modelo



MODELOS DE CLASIFICACIÓN

Métricas de los Modelos Aplicados



NO PASAR POR ALTO NINGÚN
CASO **POSITIVO** (EVITAR FALSOS
NEGATIVOS)



MODELO ORIENTADO A UNA **FASE**
DE TRIAJE

MODELOS DE CLASIFICACIÓN

Métricas de los Modelos Aplicados

RANDOM FOREST				
	Precision	Recall	F1-Score	Support
Clase 0	0.88	0.52	0.65	3139
Clase 1	0.29	0.74	0.42	854
Global			0.53	3993

XG BOOST				
	Precision	Recall	F1-Score	Support
Clase 0	0.93	0.41	0.57	2089
Clase 1	0.29	0.89	0.44	573
Global			0.51	2662

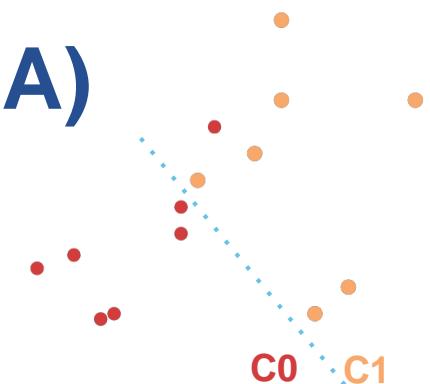
ARBOL DE DECISION				
	Precision	Recall	F1-Score	Support
Clase 0	0.86	0.69	0.77	3139
Clase 1	0.36	0.60	0.44	854
Global			0.67	3993

ANALISIS DISCRIMINANTE LINEAL (LDA)				
	Precision	Recall	F1-Score	Support
Clase 0	0.92	0.41	0.56	2089
Clase 1	0.29	0.88	0.43	573
Global			0.51	2662

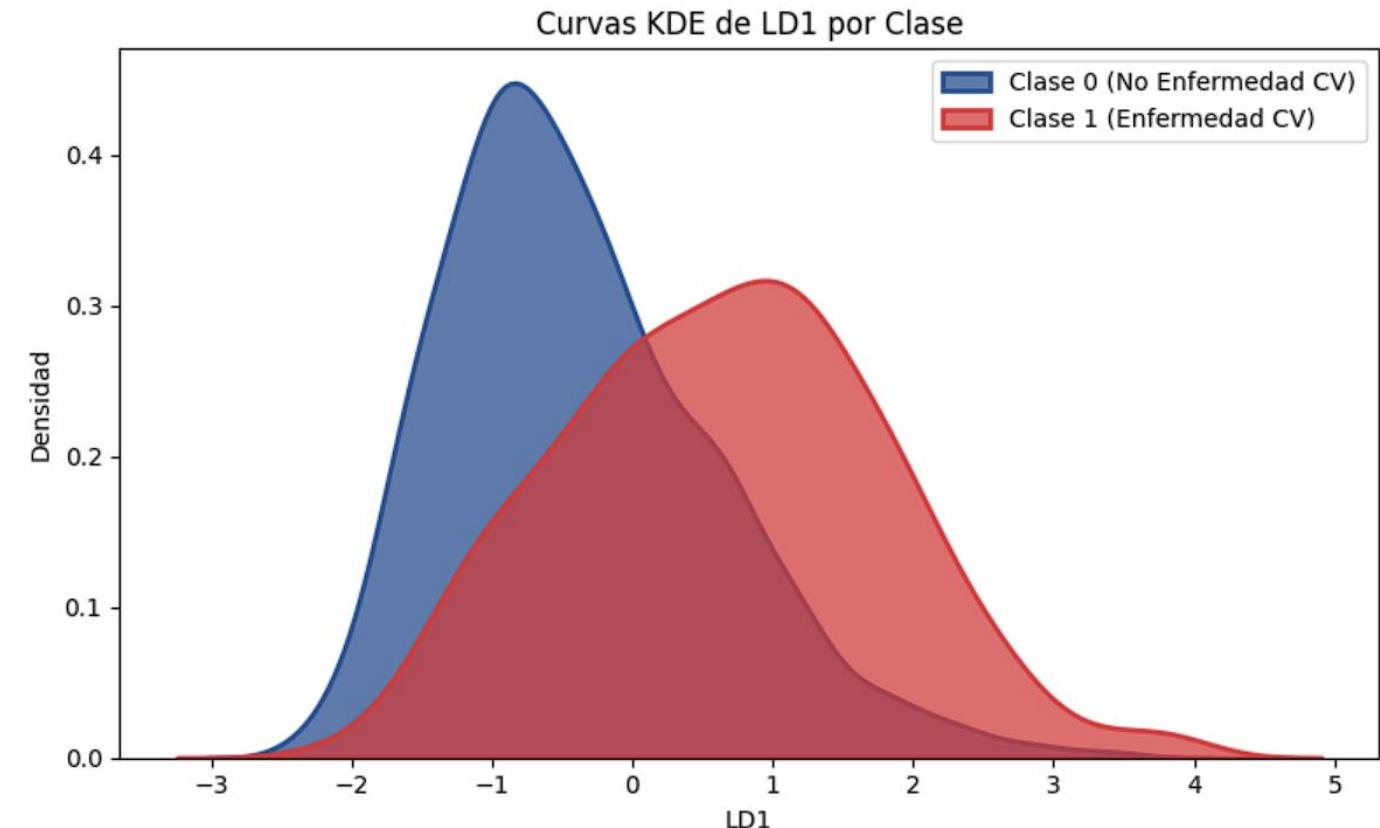
ANÁLISIS DISCRIMINANTE LINEAL (LDA)

Reducción de las Variables

$$LD(x) = b + w_1x_1 + w_2x_2 + \dots + w_px_p,$$



13308	Registros
66	Variables
	coef abs < 0.1
	13308
56	Variables



ANÁLISIS DISCRIMINANTE LINEAL (LDA)

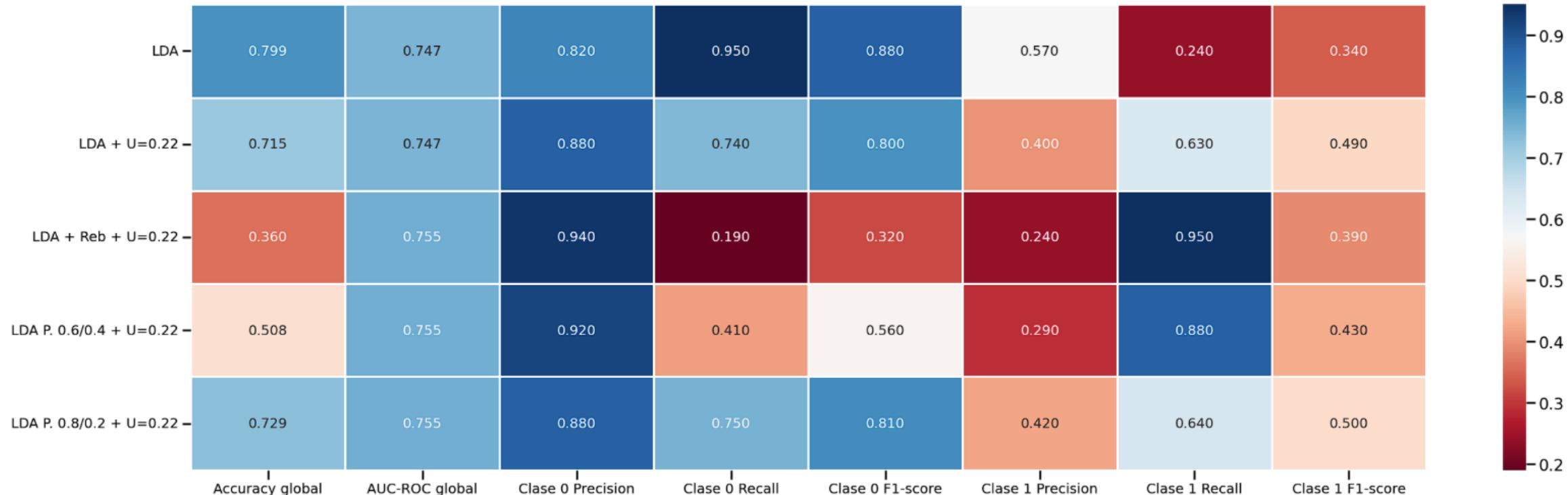
Mejoras en el Modelo

1. Ajuste de Hiperparametros
(Umbral)

2. Rebalanceo de las Clases
(Subsampleo)

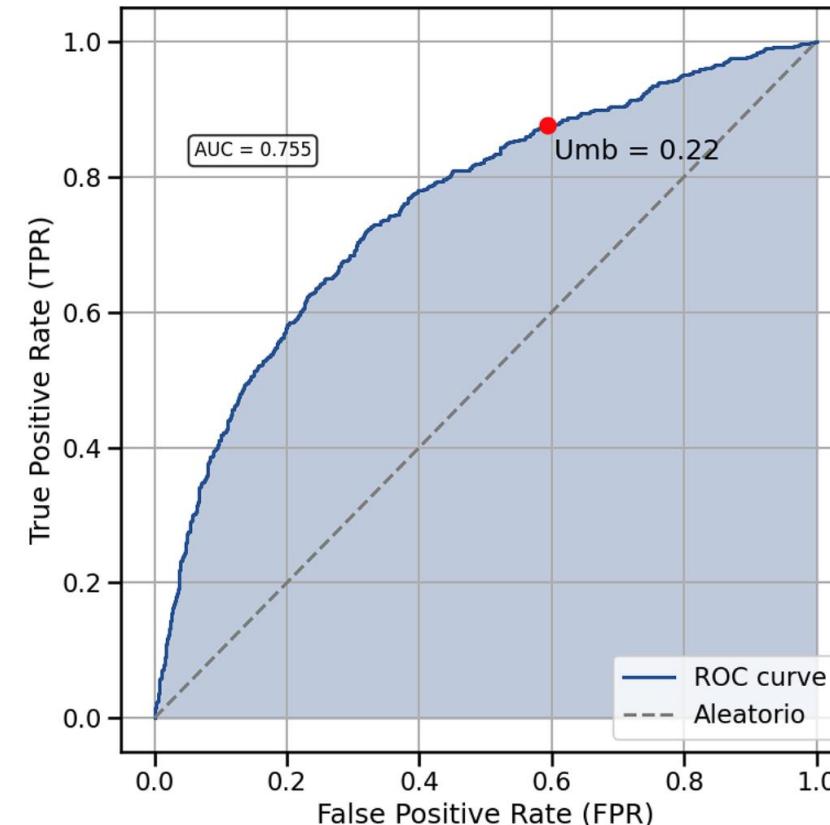
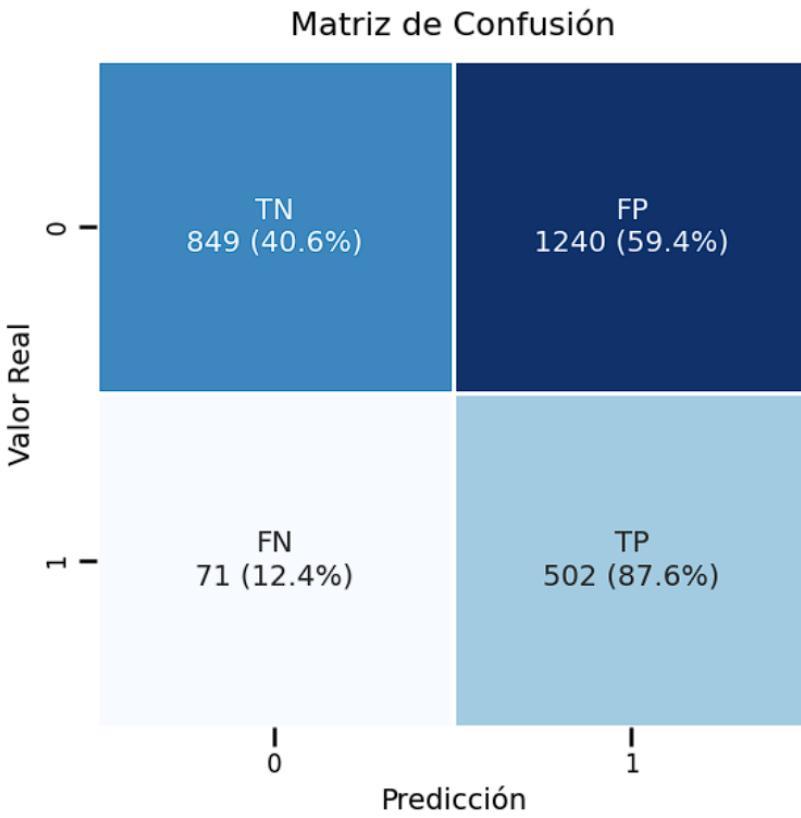
3. Ajuste de Priors

Comparación de configuraciones de LDA vs Métricas



ANÁLISIS DISCRIMINANTE LINEAL (LDA)

Métricas del Modelo Mejorado

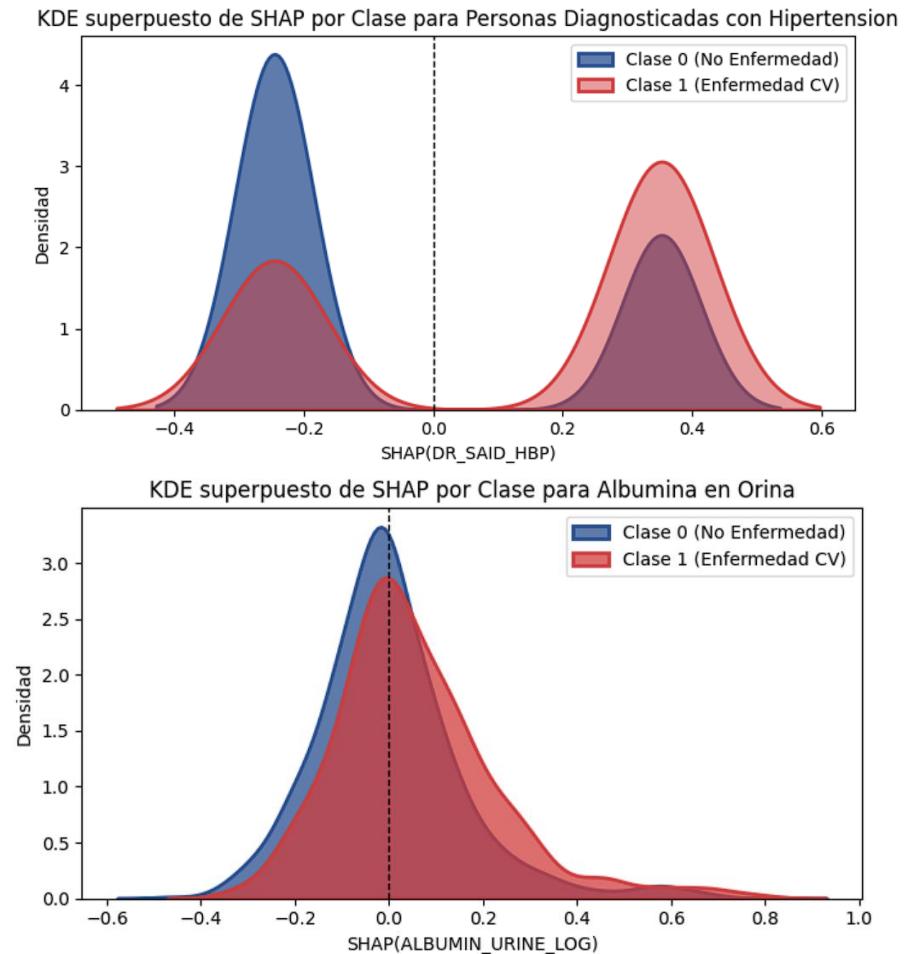
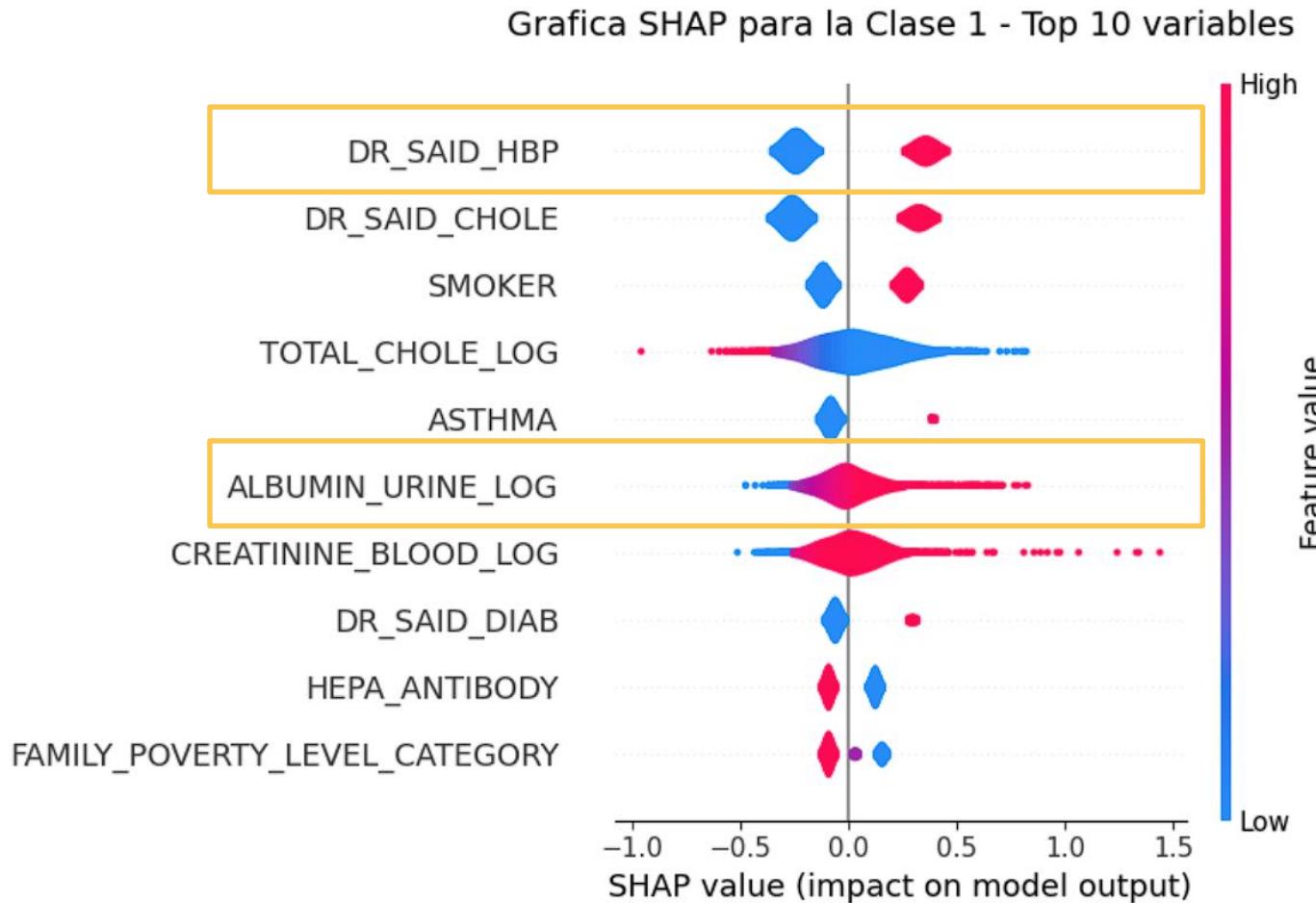


Trade-off
($U=0.22$)

Mas sensibilidad (más TP)
Menos especificidad (más FP)

ANÁLISIS DISCRIMINANTE LINEAL (LDA)

Variables Relevantes para la Clasificación de Pacientes



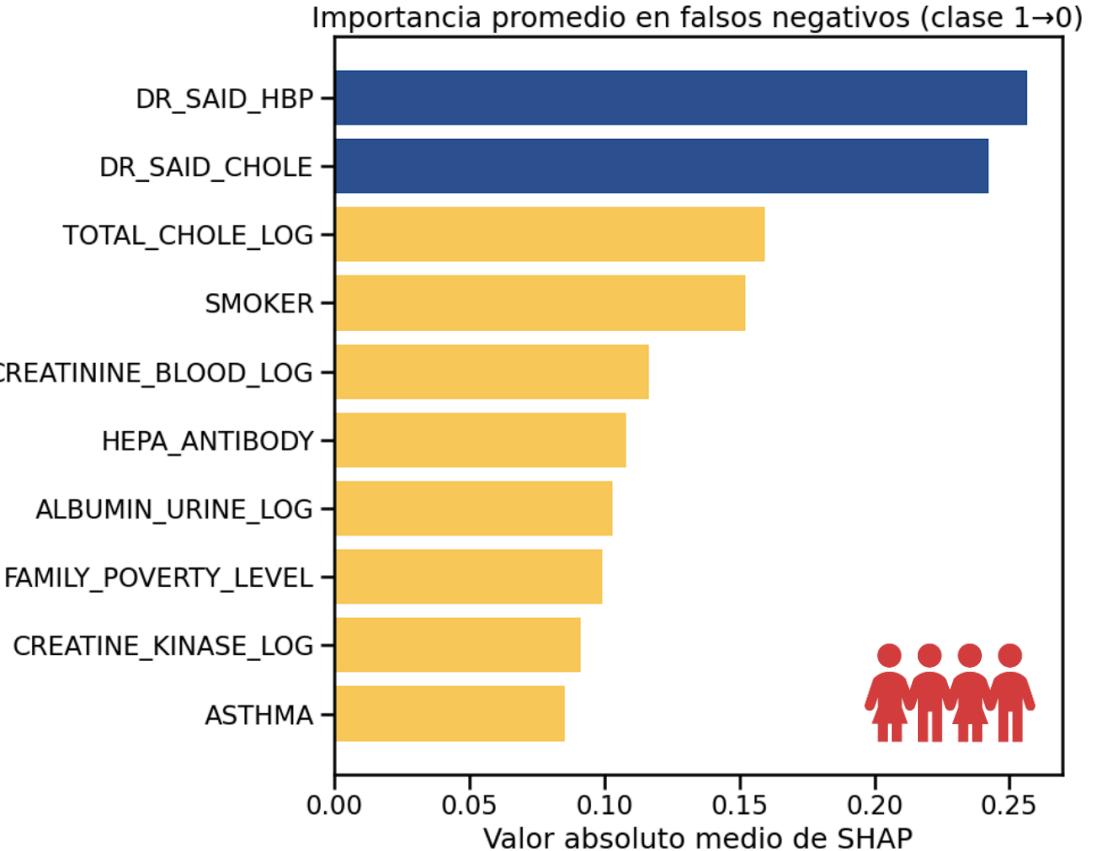
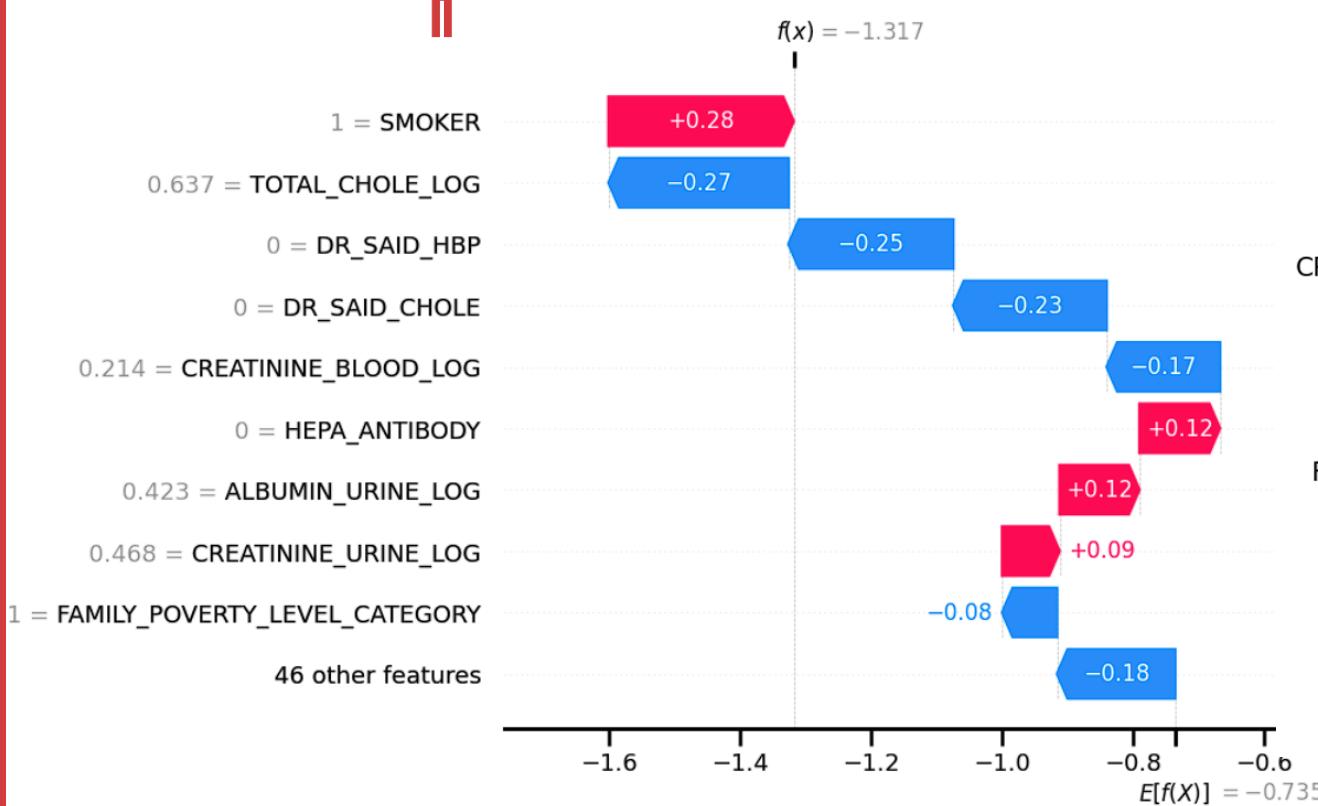
ANÁLISIS DISCRIMINANTE LINEAL (LDA)

Explicabilidad de los Falsos Negativos

71 FN



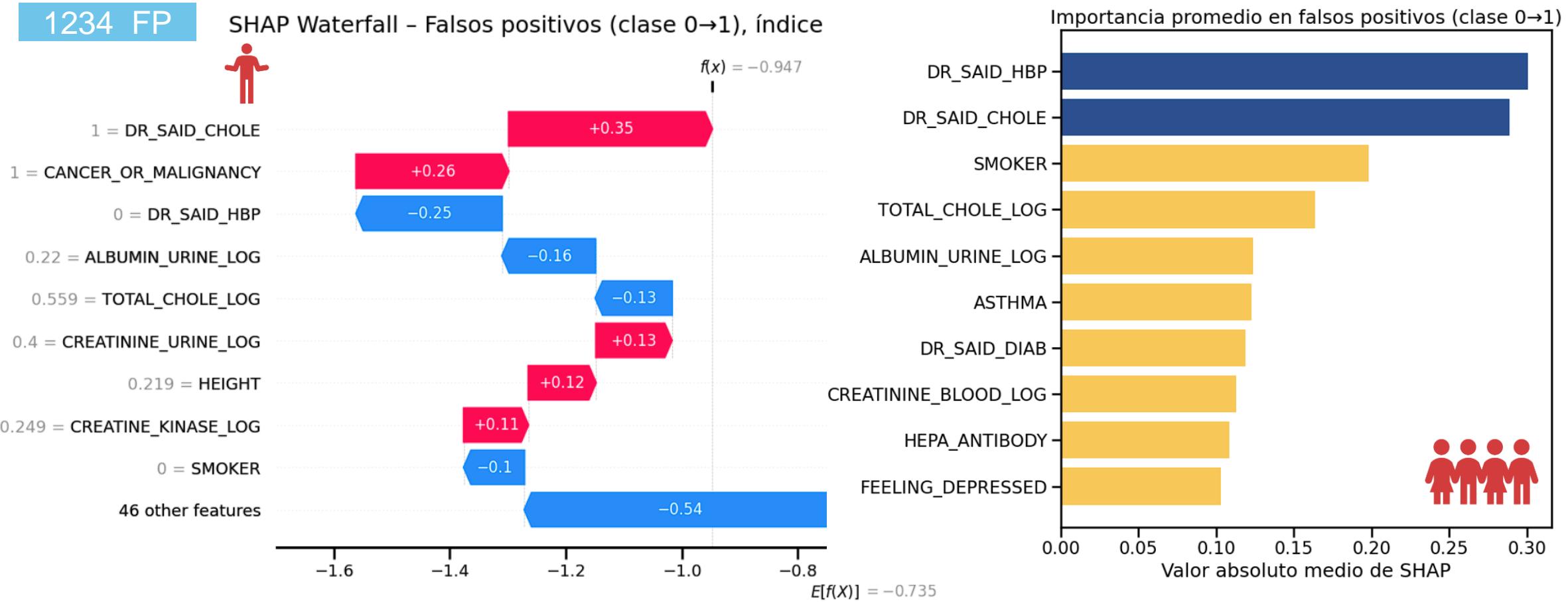
SHAP Waterfall – Fila explicada: 241



No tener hipertension o colesterol alto no implica que no se puedan tener enfermedades CV

ANÁLISIS DISCRIMINANTE LINEAL (LDA)

Explicabilidad de los Falsos Positivos



Tener hipertension o colesterol alto no implica que se tengan enfermedades CV

XGBoost

Mejoras del Modelo

XG BOOST - SOBRE MUESTREO

	Precision	Recall	F1-Score	Support
0	0.84	0.80	0.82	3139
1	0.37	0.42	0.39	854
Accuracy			0.72	3993

XG BOOST - SUB MUESTREO

	Precision	Recall	F1-Score	Support
0	0.88	0.67	0.76	3139
1	0.36	0.67	0.47	854
Accuracy			0.67	3993

XG BOOST - SMOTETOMEK

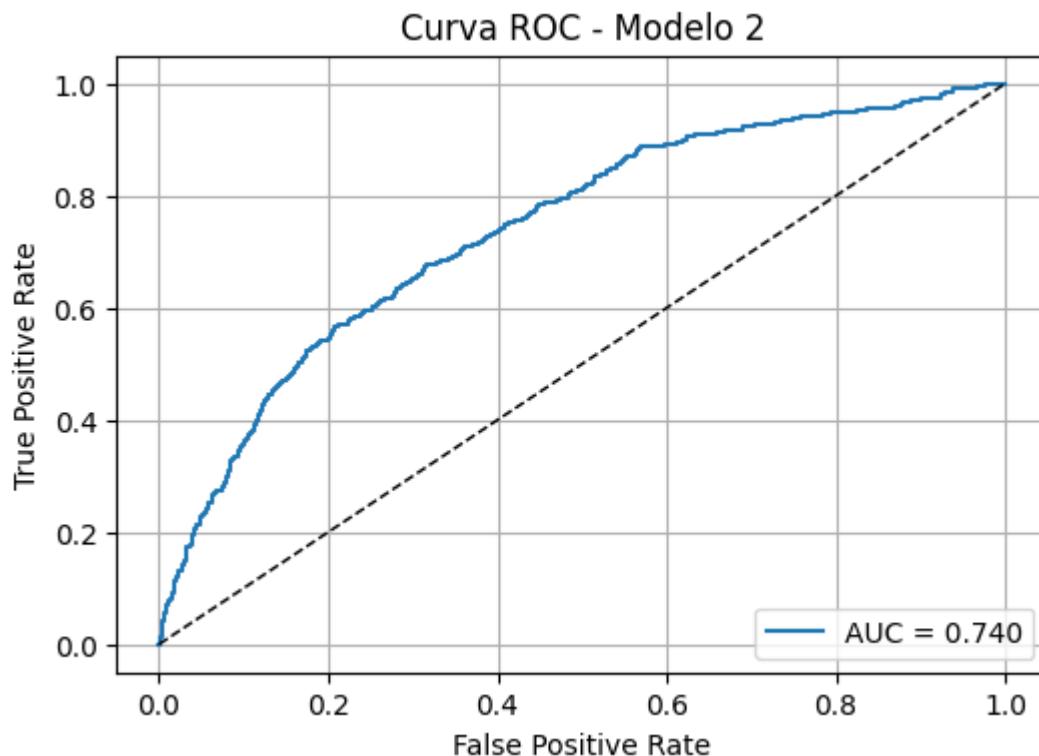
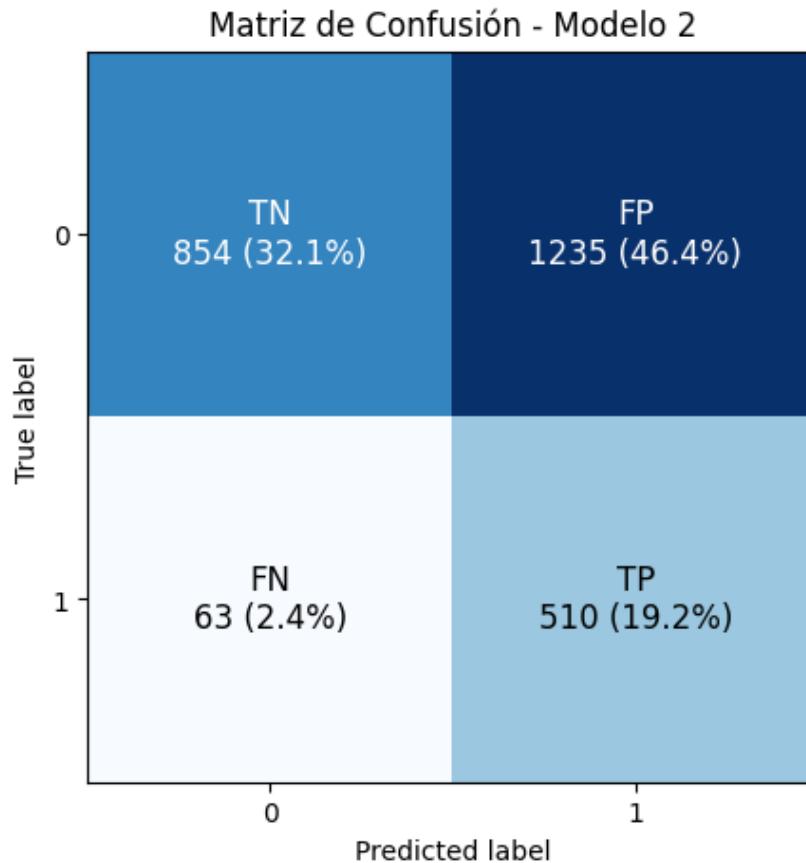
	Precision	Recall	F1-Score	Support
0	0.86	0.81	0.83	2089
1	0.43	0.53	0.48	573
Accuracy			0.75	2662

XG BOOST - SMOTETOMEK

	Precision	Recall	F1-Score	Support
0	0.93	0.41	0.57	2089
1	0.29	0.89	0.44	573
Accuracy			0.51	2662

XGBoost

Métricas del Modelo Mejorado (XGBoost + SMOTETOMEK)

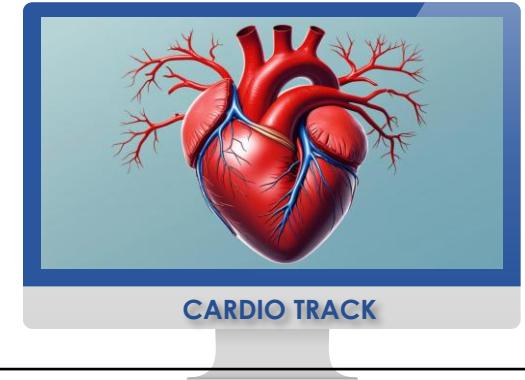


Mas sensibilidad (más TP)

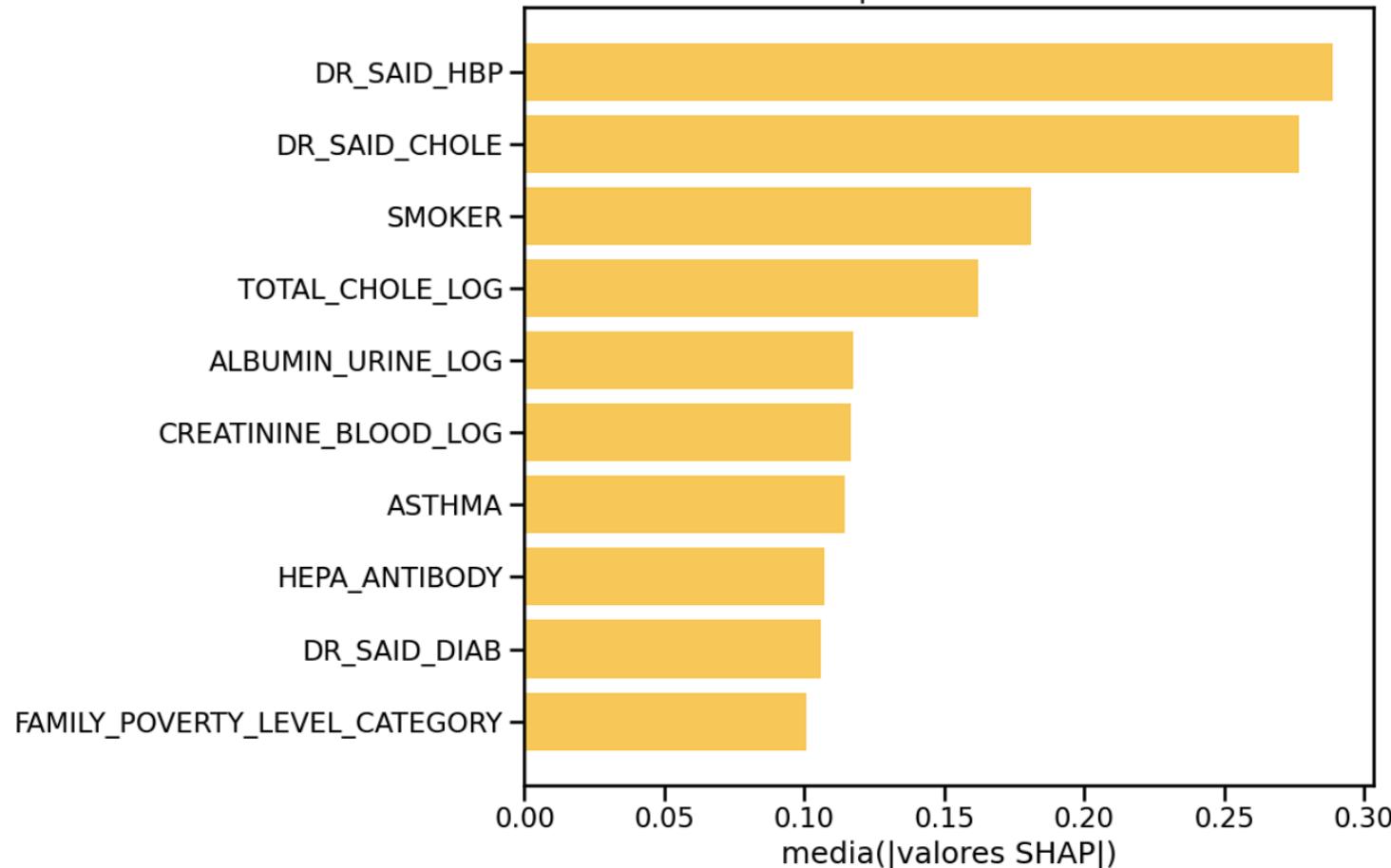
Menos especificidad (más FP)

SOFTWARE CLINICO

Cribado de Pacientes - Cardiología



🏆 TOP 10 VARIABLES PREDICTIVAS



- Diagnóstico médico de presión arterial alta (Hipertensión)*
- Diagnóstico médico de colesterol alto*
- Consumo de tabaco*
- Nivel total de colesterol en sangre*
- Nivel de albúmina en orina (indicador de función renal)*
- Nivel de creatinina en sangre (marcador de función renal)*
- Diagnóstico médico de asma*
- Presencia de anticuerpos contra la hepatitis A*
- Diagnóstico médico de diabetes*
- Categoría del nivel de pobreza del hogar*

SOFTWARE CLINICO

CardioTrack

Dr. García
Cardiólogo

Paciente: Ana Martínez

Panel Principal

Presión Arterial **120/80 mmHg** Normal

Frecuencia Cardíaca **72 bpm** Óptima

Colesterol **185 mg/dL** Deseable

Riesgo CV **Bajo** Favorable

Peso **68 kg** Normal

IMC **23.5 kg/m²** Saludable

Glucosa **95 mg/dL** Normal

Factores de Riesgo

- Fumador: No
- Peso: Normal
- Antecedentes Familiares: Hipertensión
- Consumo de Alcohol: Ocasional

Condiciones Médicas

- Asma: No
- Cáncer: No
- Tumor: No
- Hipertensión: Leve
- Enfermedades Cardiovasculares: No

Próximas Citas

- Revisión Cardiológica Dr. García 20/06/2023 - 10:30
- Análisis de Sangre Lab. Central 25/06/2023 - 08:15
- Consulta Nutrición Dra. Sánchez 02/07/2023 - 16:00

Medicación Actual

- Enalapril 5mg 1 comprimido diario por la mañana
- Aspirina 100mg 1 comprimido diario con la comida
- Atorvastatina 20mg 1 comprimido diario por la noche

Configuración

+

CONCLUSIONES DEL ESTUDIO

- El **desbalanceo** de clases ha sido el principal reto del estudio.
- Debido a las limitaciones de los datos disponibles, el sistema se ha diseñado como una **herramienta de cribado inicial**.
- El modelo prioriza la **detección temprana** sobre la precisión absoluta.





GRACIAS

