MAGISTER EN DATA SCIENCE 2026

Predicción temprana de progresión a riesgo cardiovascular para priorización preventiva en APS rural

MACHINE LEARNING II

INTEGRANTES: Claudio Cárdenas, Evelyn Sánchez

¿¿Qué revisaremos?

Contexto y problema

Establecer el contexto de las enfermedades cardiovasculares en Chile

Datos

Recopilar y organizar datos relevantes

EDA

Realizar análisis exploratorio de datos

Pipeline de preprocesamiento

Preparar los datos para el modelado

Modelos evaluados

Evaluar varios modelos de aprendizaje automático

Comparación final

Comparar el rendimiento de los modelos

Modelo seleccionado

Elegir el modelo más adecuado

Interpretabilidad

Entender las predicciones del modelo

Limitaciones

Reconocer las limitaciones del modelo

Despliegue

Implementar el modelo en un entorno del mundo real

Conclusiones

Resumir los hallazgos y recomendaciones

Contexto y problema

El PSCV incorpora solo a quienes ya cumplen criterios diagnósticos.

Sin embargo:
•Muchas personas presentan **alteraciones subclínicas progresivas**.
•No cumplen criterios PSCV → **no entran al programa**.
•Pueden progresar hacia HTA, DM2 o dislipidemia en meses/años.
•Esto genera **ingresos tardíos**, mayor costo y peor pronóstico.

Brecha real: anticipar quiénes van a progresar para priorizar intervenciones

Pregunta de investigación:
¿Es posible predecir la progresión hacia criterios de ingreso al Programa de Salud Cardiovascular utilizando modelos de aprendizaje supervisado basados en datos clínicos rutinarios de APS rural?

Tratamiento del tema:

Aumento de las muertes cardiovasculares en Chile

Sistema sanitario

El sistema sanitario está sobrecargado, especialmente la APS

Problemas rurales

El acceso limitado dificulta la prevención y la detección

Alta prevalencia

La hipertensión, la obesidad y la diabetes aumentan el riesgo

Revisión de la Literatura: análisis bibliográfico

Eje temático	Referencias clave	Principales aportes	Relevancia para el estudio
Carga de enfermedad cardiovascular y contexto chileno	DEIS (2024); ENS (2023)	Evidencian que las ECV son la principal causa de muerte en Chile y muestran un aumento sostenido de obesidad, hipertensión y diabetes.	Justifica la relevancia sanitaria del problema y la necesidad de fortalecer estrategias preventivas en APS.
APS rural y brechas en la gestión del riesgo cardiometabólico	WHO (2022); OECD (2023); Núñez et al. (2023)	Identifican limitaciones estructurales en APS rural: acceso, continuidad de cuidados y seguimiento de factores de riesgo.	Fundamenta el foco territorial del estudio y la necesidad de herramientas de priorización preventiva en contextos rurales.
Progresión cardiometabólica subclínica	Zhang et al. (2022); Perreault et al. (2022); Huang et al. (2021)	Demuestran que estados subclínicos (prehipertensión, prediabetes, dislipidemias leves) progresan a enfermedad clínica en plazos relativamente cortos.	Sustenta la definición del outcome del estudio como progresión hacia criterios de ingreso al PSCV.
Marco normativo del PSCV en Chile	Ministerio de Salud de Chile – PSCV (2017)	Define criterios diagnósticos y operativos para el ingreso al Programa de Salud Cardiovascular.	Permite operacionalizar la variable dependiente con base normativa y contextualizar la brecha del enfoque reactivo actual.

Revisión de la Literatura: análisis bibliográfico

Eje temático	Referencias clave	Principales aportes	Relevancia para el estudio
Machine Learning en predicción cardiovascular (estado del arte)	Rao et al. (2024)	Revisión narrativa del uso de IA y ML en cardiología; destaca modelos supervisados, boosting y deep learning, junto con la importancia de la interpretabilidad.	Fundamenta conceptualmente el uso de técnicas de aprendizaje automático para la estratificación temprana del riesgo.
Modelos empíricos de ML aplicados a enfermedad cardiovascular	Muhyi & Ata (2025)	Comparación empírica de XGBoost y ANFIS en múltiples datasets cardiovasculares; XGBoost muestra desempeño competitivo y consistente.	Justifica la elección de XGBoost como modelo principal para la predicción de progresión cardiometabólica.
ML y deep learning para predicción de riesgo a largo plazo	Weng et al. (2024)	Modelo basado en PPG y deep learning capaz de predecir eventos cardiovasculares a 10 años con desempeño comparable a scores clínicos tradicionales.	Refuerza el potencial de modelos ML para complementar la prevención cardiovascular, especialmente cuando se dispone de datos rutinarios o no convencionales.

Justificación del Estudio

Este estudio es necesario porque:

- Permite **identificar precozmente** a quienes progresarán hacia criterios PSCV.
- Mejora la **asignación preventiva de recursos** en APS rural.
- Reduce **sobrecarga futura** del PSCV.
- Usa **datos rutinarios** accesibles y de bajo costo.
- Está alineado con recomendaciones internacionales de salud poblacional.

Objetivo general

Desarrollar y evaluar un modelo predictivo que permita estimar la probabilidad de progresión hacia criterios de ingreso al Programa de Salud Cardiovascular (PSCV) en usuarios que actualmente no cumplen dichos criterios, como apoyo a la priorización preventiva, utilizando datos clínicos rutinarios de Atención Primaria de Salud rural.

Objetivos específicos

1. Preparar y depurar registros clínicos provenientes de EMPA y Control Cardiovascular en APS rural.
2. Definir una variable objetivo asociada a la progresión hacia criterios de ingreso al PSCV.
3. Entrenar y comparar modelos de aprendizaje supervisado para la estimación del riesgo cardiovascular.
4. Evaluar el desempeño de los modelos mediante métricas orientadas a contextos preventivos.
5. Analizar la interpretabilidad del modelo seleccionado mediante técnicas de explicabilidad (SHAP).
6. Proponer un esquema de priorización preventiva y uso conceptual del modelo en APS rural.

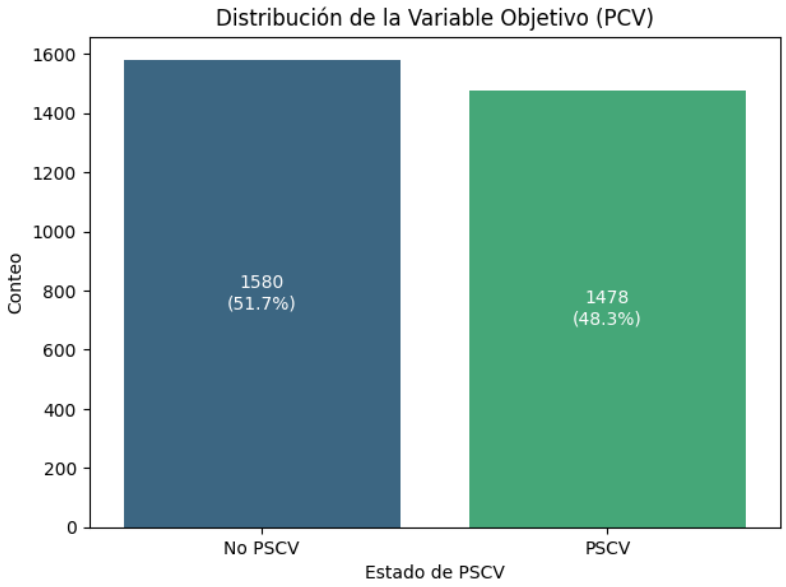
Descripción del Dataset

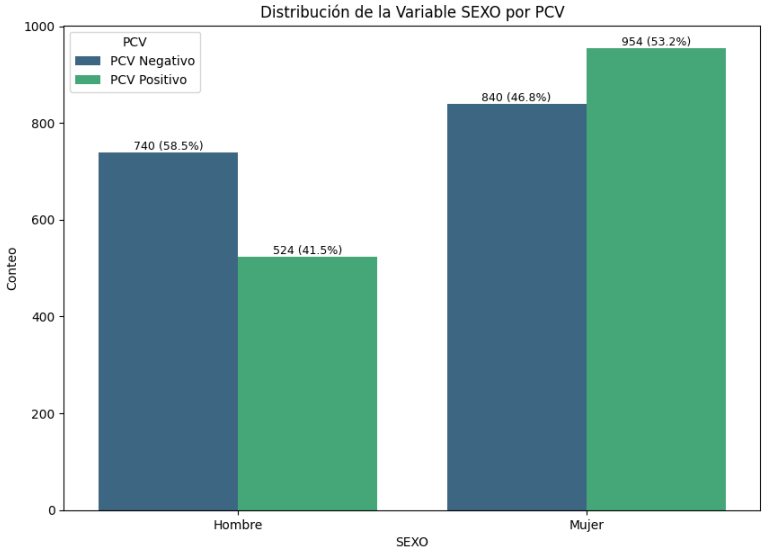
Fuente: APS Quellón
Registros: 3.058 usuarios (2021–2023)
Variables incluidas:
•Edad, Sexo, PAS / PAD, Circunferencia cintura, Glicemia, Colesterol total

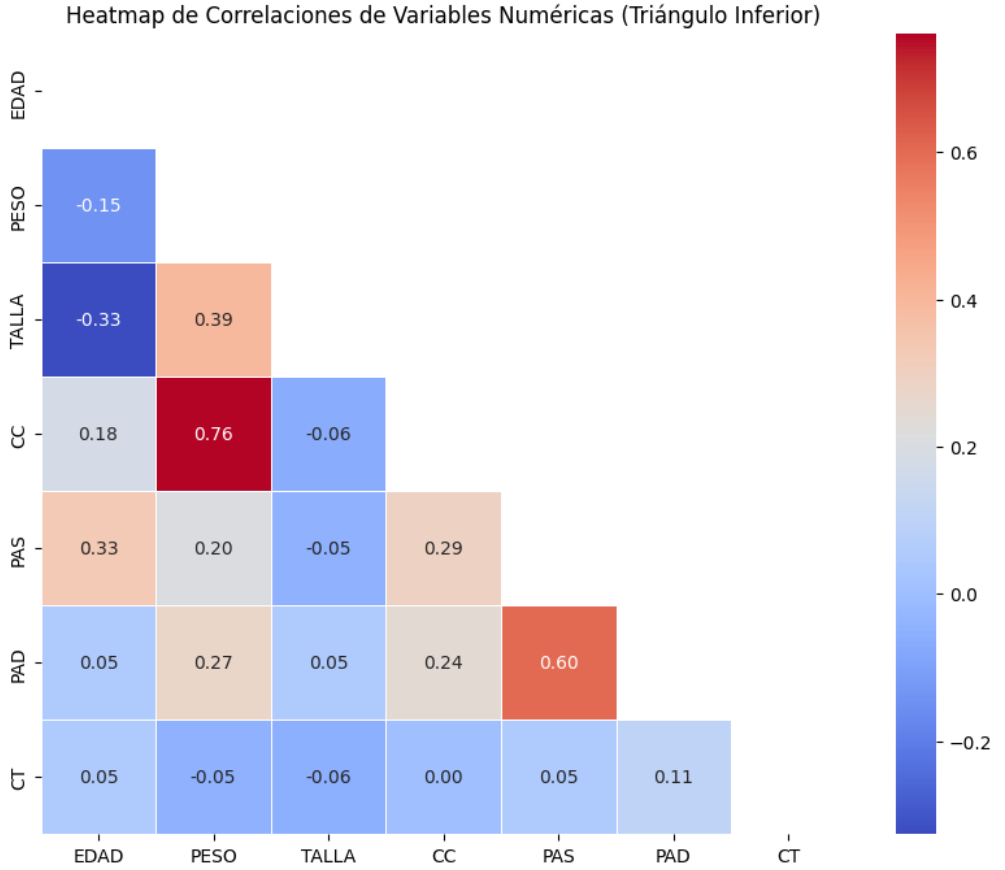
Variable objetivo:
•**Progresión PSCV** (1 = progresa, 0 = no progresa)

	PCV	SEXO	EDAD	PESO	TALLA	CC	PAS	PAD	CT
0	1	0	56	110.2	168	119.0	126	80	275.0
1	1	1	81	70.0	144	97.0	130	60	171.0
2	1	1	60	92.4	155	110.0	180	86	216.0
3	1	1	84	70.2	152	113.0	116	70	198.0
4	1	0	76	77.1	150	107.0	180	80	223.0

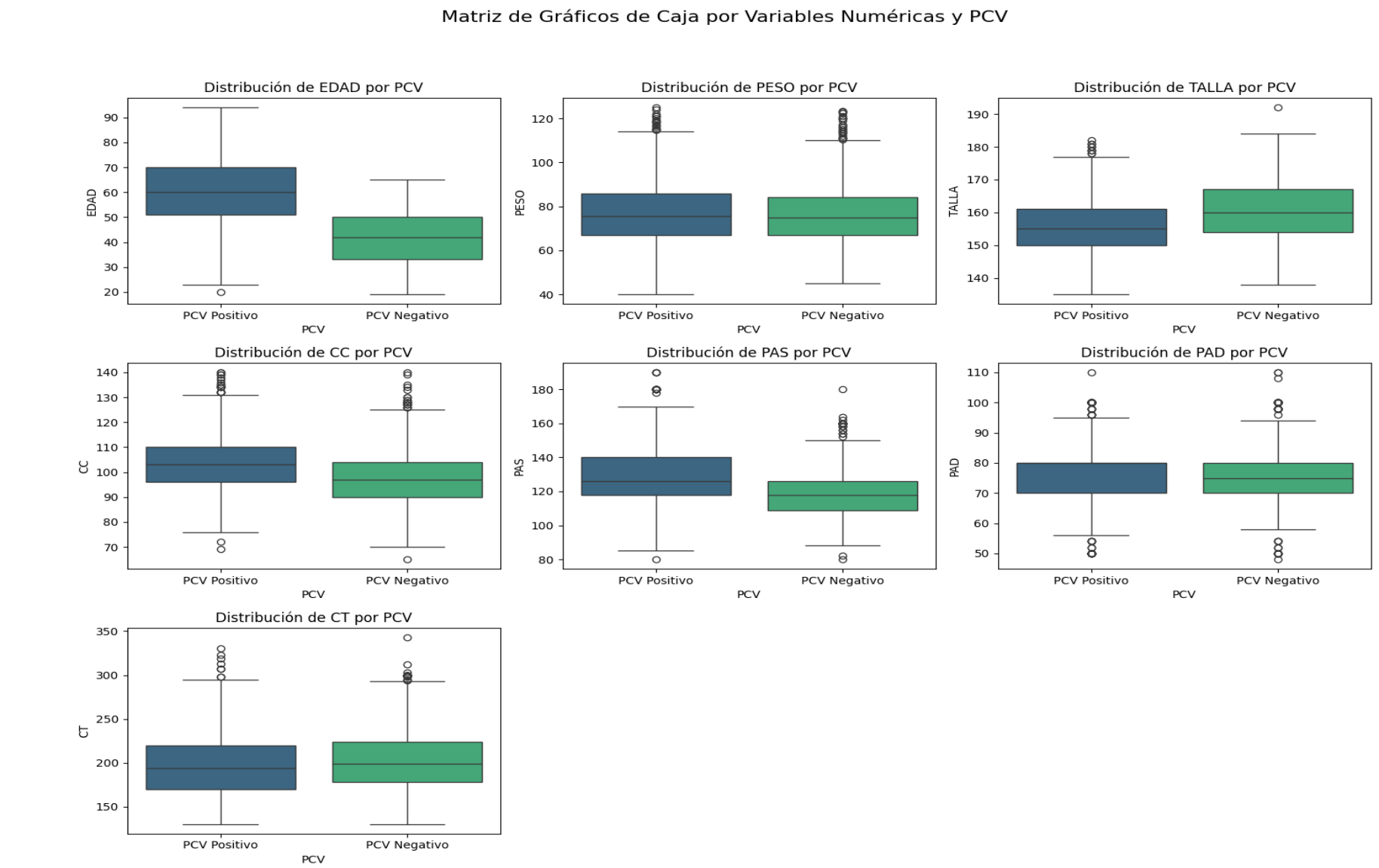
EDA: Visualización de datos



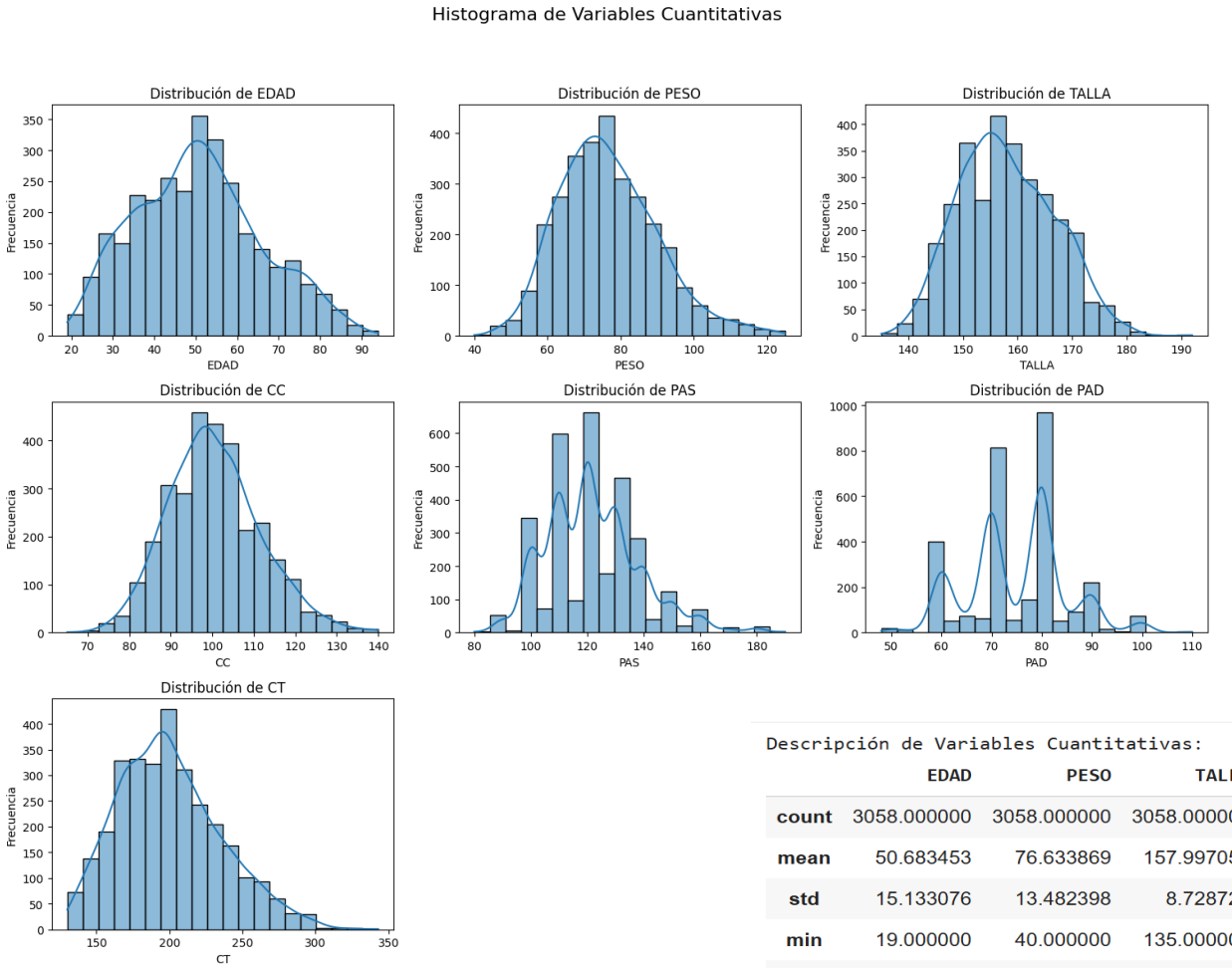




Visualización de datos



Técnicas de análisis de datos:



Descripción de Variables Cuantitativas:

	EDAD	PESO	TALLA	CC	PAS	PAD	CT
count	3058.000000	3058.000000	3058.000000	3058.000000	3058.000000	3058.000000	3058.000000
mean	50.683453	76.633869	157.997057	100.423708	121.773054	74.954545	200.190320
std	15.133076	13.482398	8.728728	11.166665	16.367233	9.872042	35.354991
min	19.000000	40.000000	135.000000	65.000000	80.000000	48.000000	130.000000
25%	39.000000	67.000000	151.000000	93.000000	110.000000	70.000000	174.000000
50%	50.000000	75.000000	157.000000	100.000000	120.000000	78.000000	196.000000
75%	60.000000	85.000000	164.000000	107.000000	130.000000	80.000000	222.000000
max	94.000000	125.000000	192.000000	140.000000	190.000000	110.000000	343.000000

Procesamiento de Datos

Pipeline de Preprocesamiento

1. Integración y Depuración

DataFrame

	Edad	Sexo	Sex
	72		
	78		
	37		

- Registros EMPA y Control CV
- Eliminación de:
  - NaN
  - Duplicados
  - Columna innecesaria

2. Tratamiento de Valores Atipicos

- Método IQR

	Edad	Sexo
	50	-1.1
	62	0.0
	71	1.3

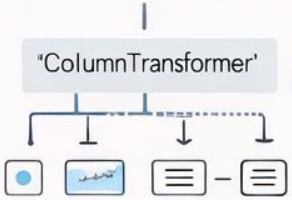
- Eliminación de outliers

3. Escalamiento y Codificación

- **Escalamiento Z-score**
- Eliminación de outliers extremos

Edad	Sexo	Sexo_M
50	-1.1	1.1
62	0.0	1.3

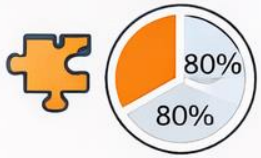
4. Pipeline de preprocesamiento



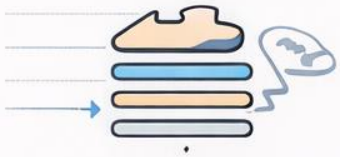
- **Paso categórico:**
- OneHotEncoder (drop='first')

Edad	Sexo_F	Sexo_M
H	1	0
M	0	1

5. División Entrenamiento/Prueba



- 80 % entrenamiento
- 20 % prueba



Pipeline de preprocesamiento

- **Paso numérico:** StandardScaler()
- IQR
- **Paso categórico:** OneHôtEncder(drop='first')

Modelos evaluados

Modelos lineales

- ✔ Reg. Logística base
- ✔ Reg. Logística L1 / L2
- → Interpretabilidad y referencia clínica

Modelos no lineales

- ✔ Reg. Logística polinómica
- ✔ KNN
- → Exploración de mayor complejidad y recall

Modelo ensemble

- ✔ XGBoost
- → Mejor equilíbrio global
- → Modelo seleccionado

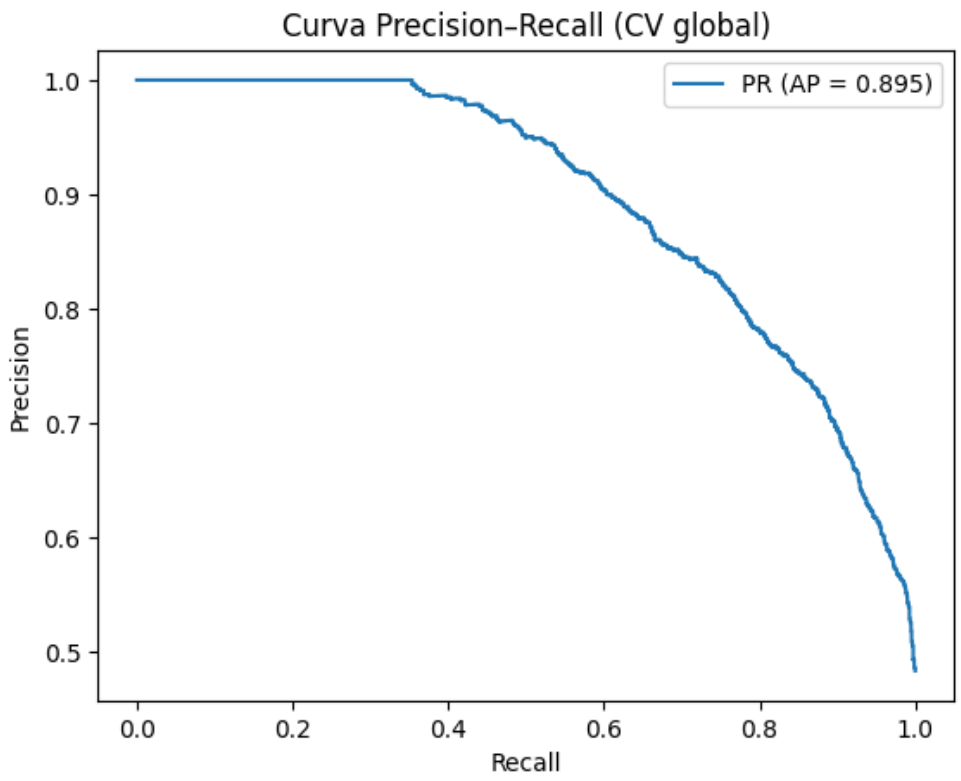
No todos los modelos evaluados están orientados a su implementación operativa algunos se utilizaron con fines exploratorios y comparativos.

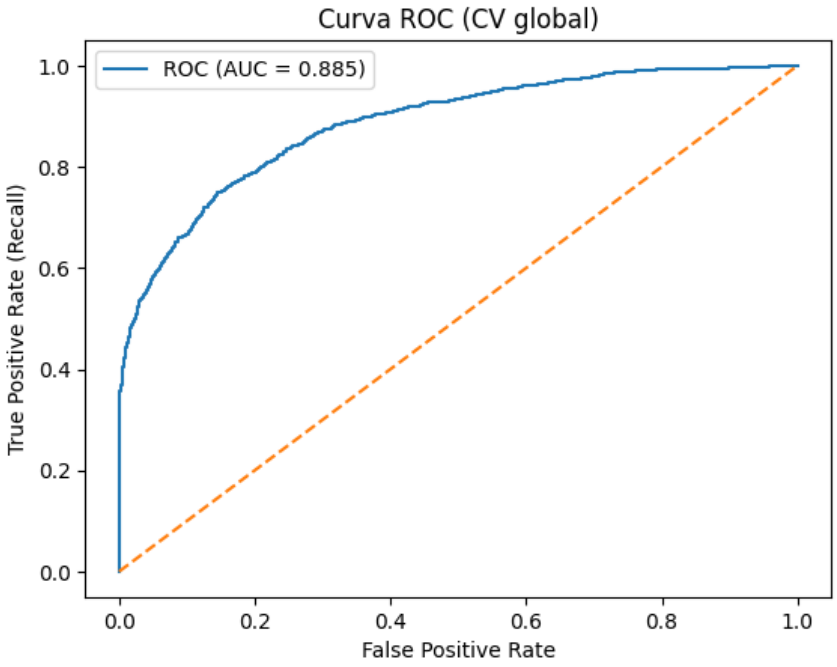
Comparación final

Tabla comparativa de desempeño de modelos

Modelo	AUC-ROC	PR-AUC	Precision (PCV=1)	Recall (PCV=1)	F1-Score (PCV=1)
XGBoost (Optimizado)	0.884815	0.894811	0.817055	0.758457	0.786667
Regresión Logística L1 (Optimizada)	0.885986	0.894478	0.825797	0.753721	0.788115
Regresión Logística con Polinomios	0.885283	0.893759	0.822664	0.756428	0.788157
Regresión Logística L2 (Optimizada)	0.885283	0.893759	0.822664	0.756428	0.788157
SVM (Optimizado)	0.882126	0.890738	0.808646	0.771989	0.789893
Regresión Logística Base	0.875316	0.887069	0.805072	0.751691	0.777467
KNN (Optimizado)	0.864444	0.865008	0.810706	0.707037	0.755331

Modelo seleccionado: Modelo XGBoost





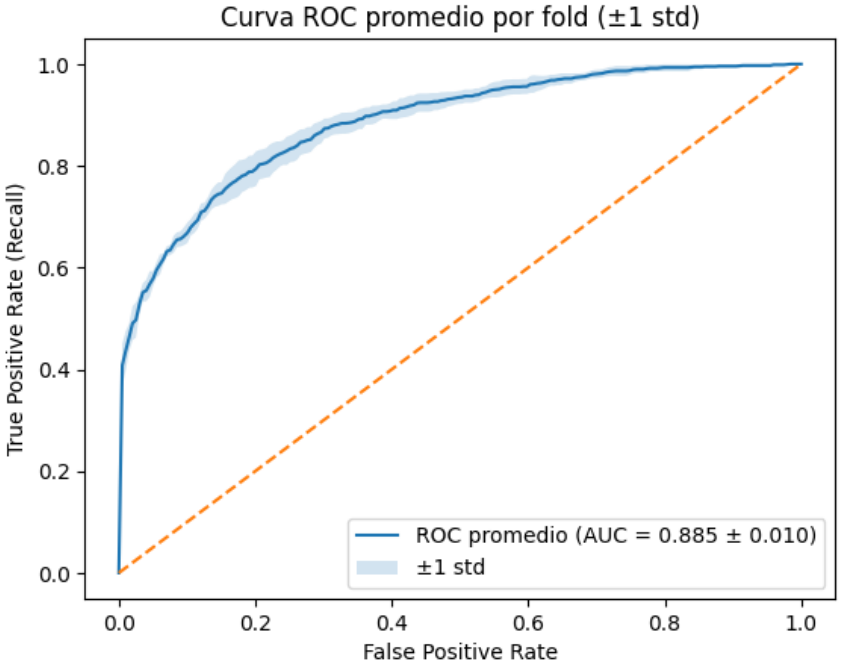
Matriz de confusión global

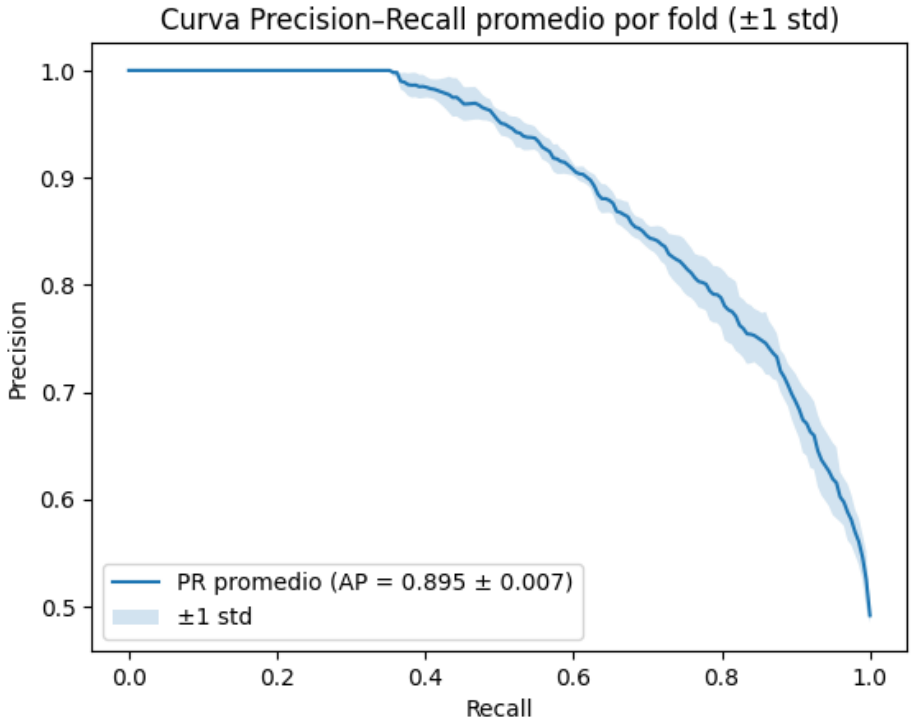
```
[[1329  251]
 [ 357 1121]]
```

Reporte de clasificación:

	precision	recall	f1-score	support
0	0.788	0.841	0.814	1580
1	0.817	0.758	0.787	1478
accuracy			0.801	3058
macro avg	0.803	0.800	0.800	3058
weighted avg	0.802	0.801	0.801	3058

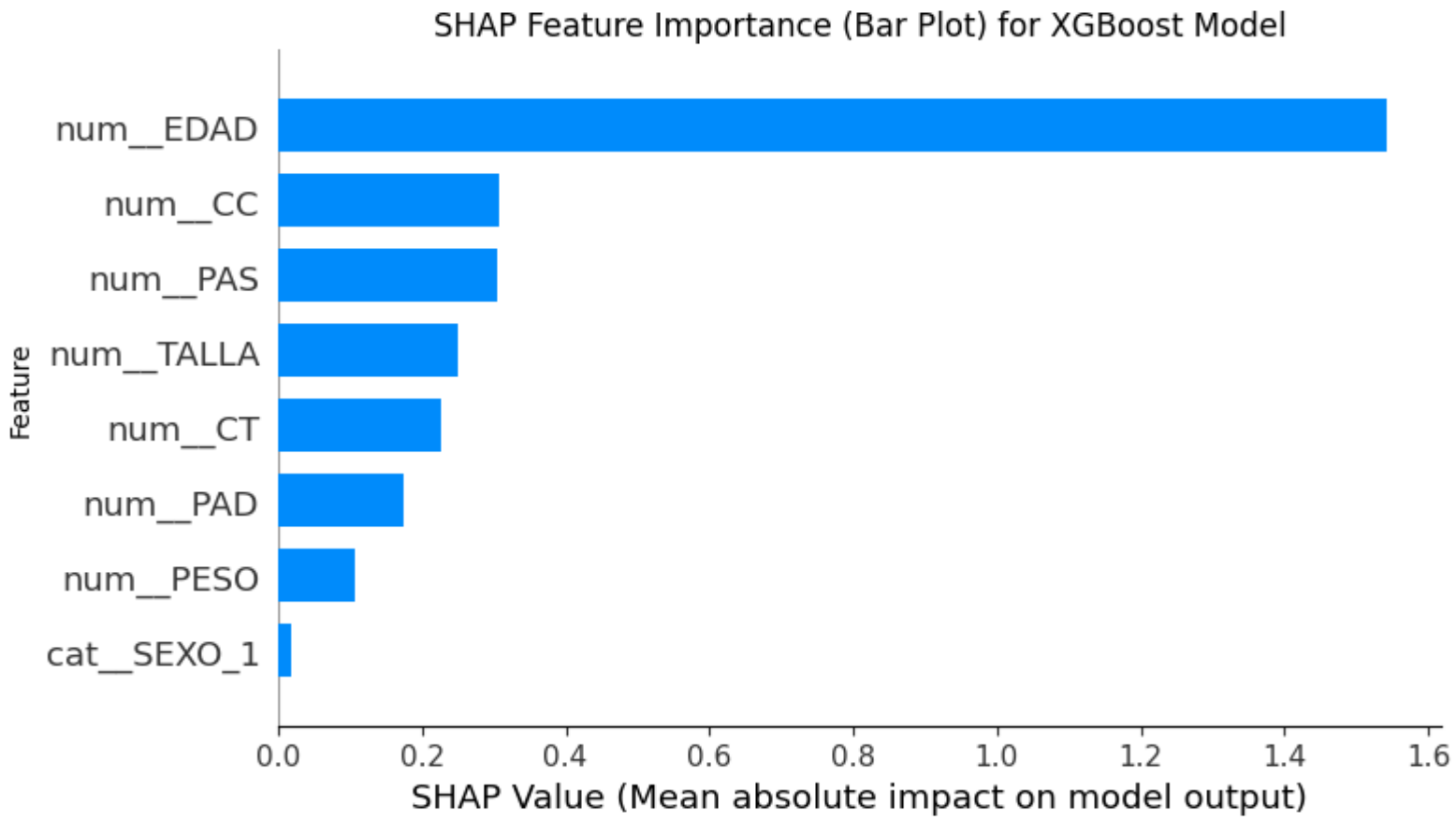
Modelo seleccionado: Modelo XGBoost





Interpretabilidad

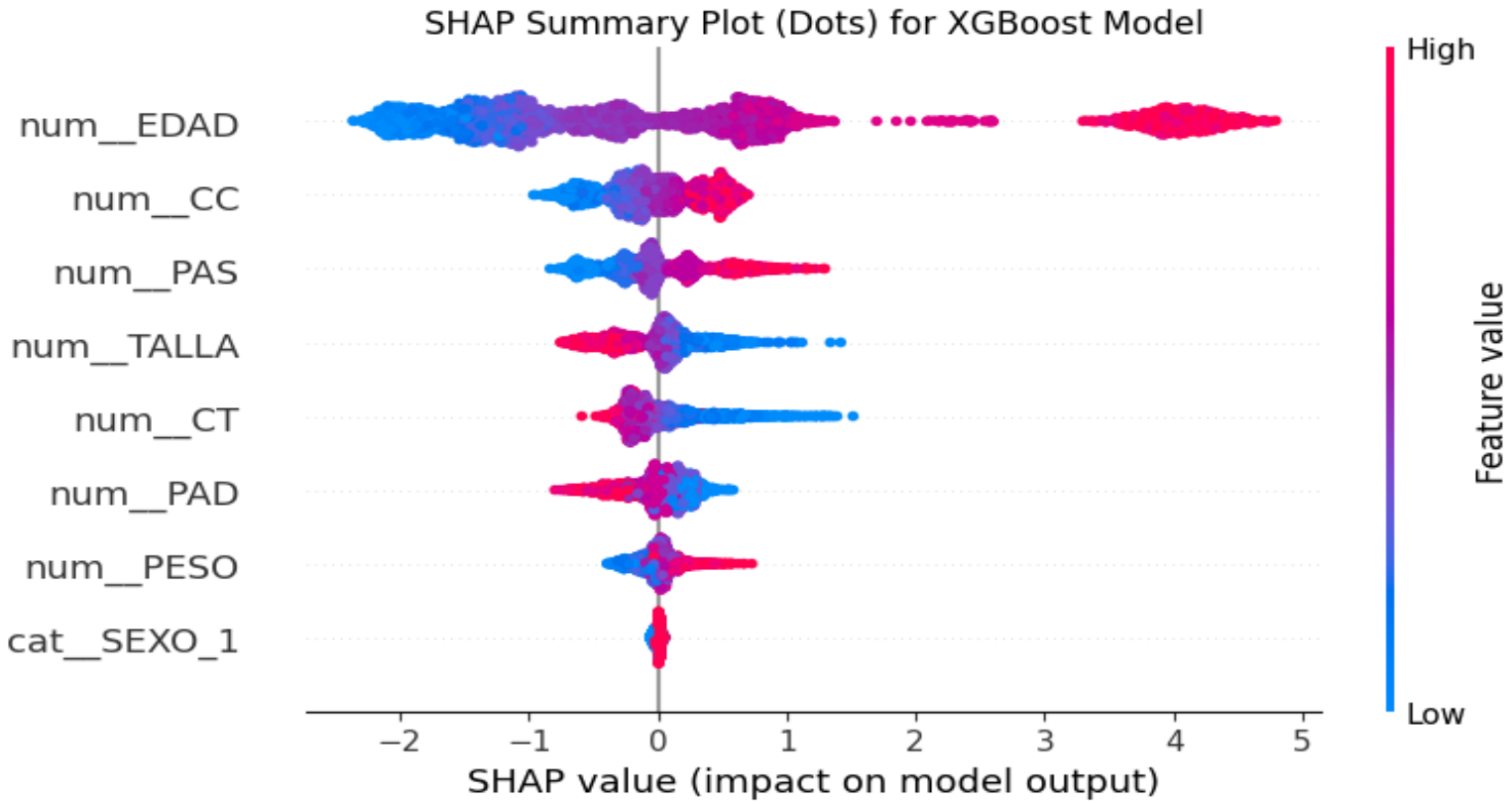
Importancia de las Características



El modelo asigna mayor riesgo a valores elevados de edad, CC y PAS, en concordancia con la evidencia clínica.

Interpretabilidad

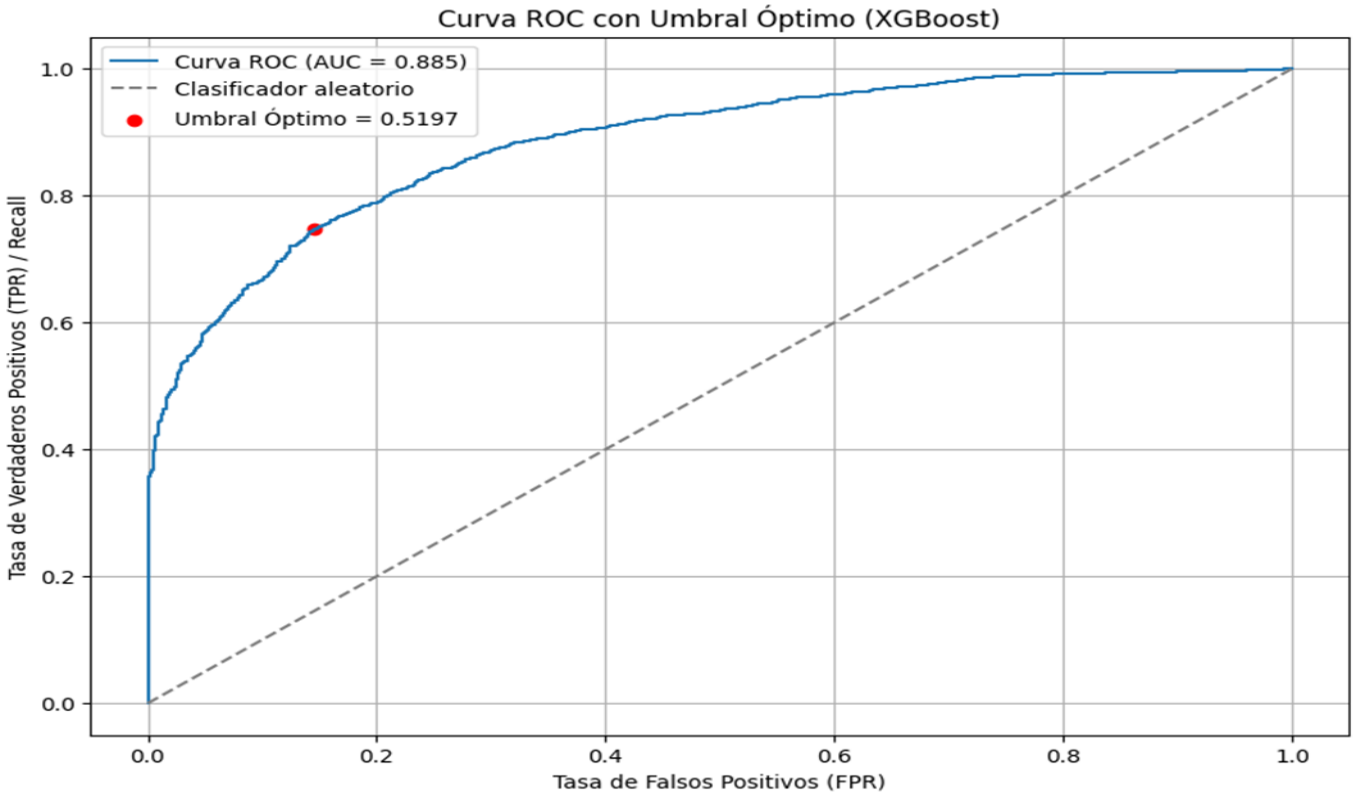
Importancia de las Características



El eje horizontal muestra el impacto de cada variable sobre la predicción del riesgo. Valores altos de edad, CC y PAS desplazan la predicción hacia mayor riesgo, mientras que valores bajos la reducen.
La dispersión observada indica variabilidad interindividual en el efecto de estas variables.

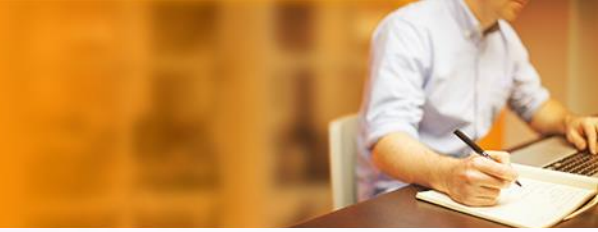
Interpretabilidad

Determinación del punto de corte (umbral de clasificación) Para el modelo XGBoost optimizado



Este valor se propone como punto de corte operativo para clasificar a las personas en alto riesgo de progresión a PSCV, balanceando sensibilidad y especificidad en un contexto de priorización preventiva en APS.

Resumen



Etapa	Decisión clave	Justificación
Definición del problema	Enfoque en progresión a PSCV (PCV=1)	Permite priorización preventiva antes del ingreso formal al programa
Prepro-cesamiento	Pipeline con escalamiento y codificación	Evita fuga de información y asegura replicabilidad
Validación	Validación cruzada estratificada (5-fold)	Mantiene balance de clases y estabilidad del desempeño
Métrica principal	Uso de PR-AUC	Más adecuada en contexto de riesgo y clases desbalanceadas
Selección de modelo	XGBoost optimizado	Mejor equilibrio entre recall, precisión y utilidad operativa
Interpreta-bilidad	SHAP	Transparencia clínica y control de sesgos
Umbral Uso del modelo	Punto de corte con índice de Youden	Balancea sensibilidad y espécificidad para APS

Limitaciones

Datos históricos y locales

El modelo fue entrenado con registros de **APS rural de Quellón**, lo que puede limitar su generalización a otros contextos poblacionales.

Variables disponibles

No incorpora hábitos de vida (**alimentación**, actividad fisica, tabaco), adherencia terapéutica ni uso de fármacos, factores relevantes en el riesgo cardiovascular.

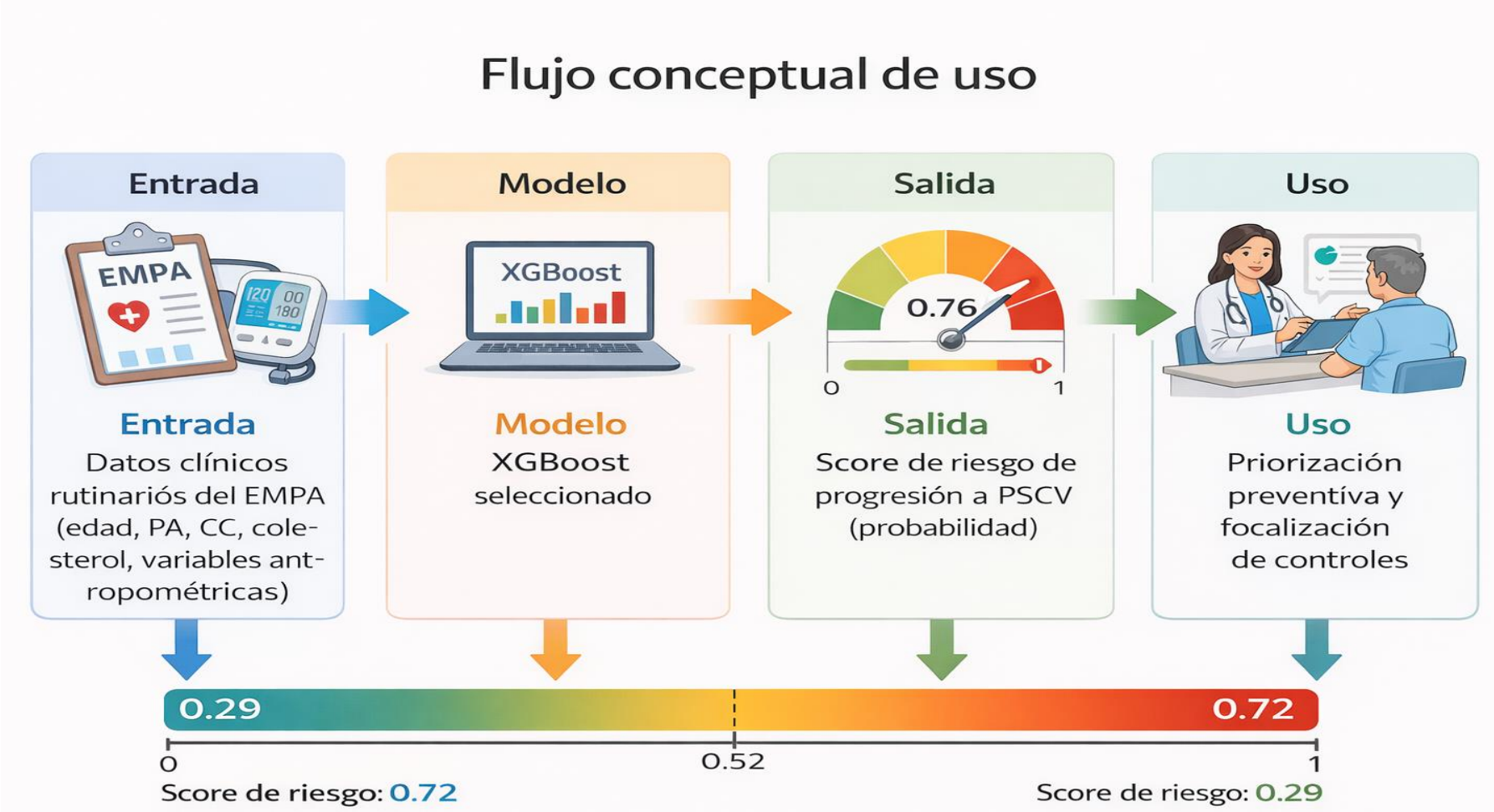
Posible subregistro y falsos negativos

Usuarios sin control activo podrían ser clasificados como de bajo riesgo pese a presentar riesgo cardiovascular no detectado.

Riesgo ≠ diagnóstico

El modelo entrega una **estimación probabilistica** y debe utilizarse solo como apoyo a la toma de decisiones clínicas.

Propuesta de despliegue del modelo en APS



Consideraciones operativas:

- Reentrenamiento periódico (p. ej., anual)
- Uso como apoyo a la decisión clínica, no diagnóstico automático

Despliegue:

Ejemplo comparativo con dos usuarios hipotéticos evaluados en el contexto de APS durante el EMPA

Usuario A — Alto riesgo

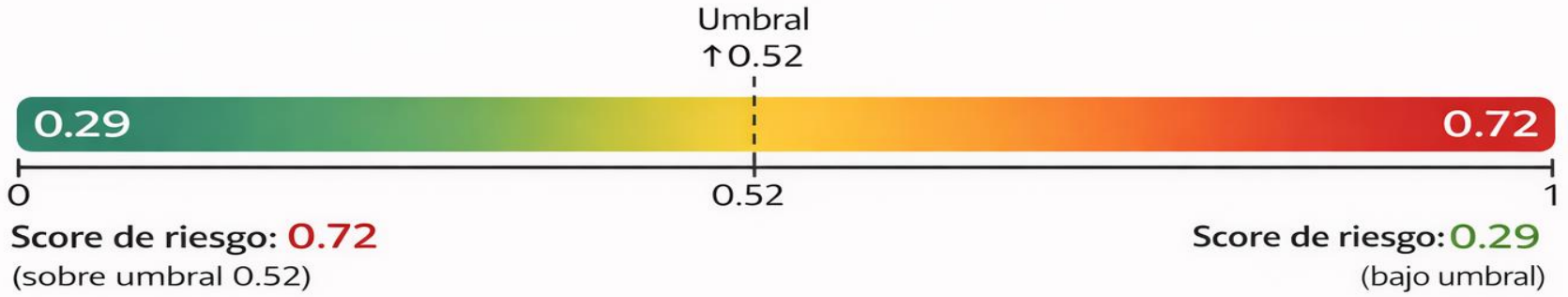
- **Edad:** 59 años
- **Sexo:** Masculino
- PAS: 150 mmHg
- PAD: 94 mmHg
- CC: 106 cm
- Colesterol total: 238 mg/dL
- IMC: 31,0 kg/m²

Score de riesgo: 0.72 ⚠
(sobre umbral 0.52)

Usuario B — Bajo riesgo

- Edad: 45 años
- Sexo: Femenino
- PAS: 122 mmHg
- PAD: 78 mmHg
- CC: 82 cm
- Colesterol total: 188 mg/dL
- IMC: 24,3 kg/m²

Score de riesgo: 0.29 ✓
(bajo umbral)



Consideraciones sobre sesgos y uso responsable del modelo

Posibles sesgos en los datos

El modelo se entrena con registros clínicos reales de APS, que pueden reflejar sesgos de selección, subregistro o características propias de la población atendida.

Estrategias de mitigación

Se utllizó validación cruzada **estratificada** y **métricas robustas** (PR-AUC) adecuadas para clases potencialmente desbalanceadas.

Transparencia del modelo

Técnicas de interpretabilidad (SHAP) permiten **identificar variables influyentes** y detectar patrones inesperados o potencialmente sesgados.

Uso responsable

El modelo apoya la **priorización preventiva** y no reemplaza el juicio clínico ni la decisión profesional.

Conclusiones

Es factible utilizar registros clínicos rutinarios de APS para estimar el riesgo de progresión a criterios del PSCV, sin requerir información clinica compleja ni exámenes de alto costo.

Los modelos evaluados mostraron desempeños altos y consistentes, evidenciando una adecuada capacidad **discriminativa** para la priorización preventiva.

XGBoost optimizado presentó el **mejor equilibrio** entre desempeño predictivo y utilidad operativa, resultando adecuado como **herramienta principal** de estratificación de riesgo.

La regresión logistica L1 mostró un rendimiento comparable y mayor **interpretabilidad,** posicionándose como un modelo **complementario** para respaldo clínico e institucional.

El enfoque propuesto permite **optimizar la priorización preventiva,** apoyar la planificación del PSCV y podria escalarse a otros contextos de APS, **previa validación externa.**

GRACIAS!