

# Proyecto Médico: Predicción de Enfermedad Cardíaca

Análisis Conceptual y Desarrollo con Python + Scikit-learn

---

Enrique Alexis Palacios Andrade

Lima, 16 de junio de 2025

Seminario - Minería de Datos

- **Reducción de mortalidad:** identificar a pacientes en riesgo para prevenir IAM.
- **Optimizar recursos:** camas UCI, laboratorios y quirófanos para alto riesgo.
- **Costos y eficiencia:** menos reingresos y eventos adversos.
- **Calidad de atención:** decisiones clínicas basadas en datos.

**Meta:** *Reducir en al menos 15% los reingresos en 30 días.*



- **Pregunta:** ¿Es posible predecir si un paciente tiene enfermedad cardíaca a partir de 13 variables clínicas?
- **Modelo:** Regresión logística + interpretabilidad (SHAP/LIME).
- **Salida:** score de riesgo individual + alerta.
- **Validación:** cross-validation estratificada.



- Variable objetivo: *condition*  $\rightarrow$  {0: sano, 1: enfermo}.
- Métricas: ROC-AUC, F1, Brier score, curvas de calibración.
- Costo asimétrico: prioridad al **recall** para minimizar falsos negativos.



# Naturaleza de Variables

Variable	Tipo	Descripción	Clínica
age	Numérica	Edad	Factor no modificable
chol	Numérica	Colesterol	Aterosclerosis
thalach	Numérica	FC máx.	Capacidad funcional
oldpeak	Numérica	Depresión ST	Isquemia
cp	Categórica	Dolor torácico	Valor diagnóstico
thal	Categórica	Prueba talio	Riesgo alto
slope	Ordinal	Pendiente ST	Isquemia
sex, fbs	Binaria	Sexo y glucemia	Riesgo base
ca	Ordinal	Obstrucción coronaria	Daño anatómico



- Fuente: *heart\_cleveland\_upload.csv*
- 297 registros, 14 variables, sin valores nulos.
- 68 % hombres, edad media: 54.5 años.
- Colesterol promedio: 247 mg/dL.
- 46 % con condición cardíaca positiva.



- *Boxplots* revelan outliers en *chol*, *oldpeak*.
- *Heatmap* muestra alta correlación con *condition* en:
  - *thal*, *ca*, *oldpeak*, *exang*.
- Ligera multicolinealidad entre *oldpeak* y *slope*.



- Categóricas: *cp*, *thal*, *slope* → OneHotEncoder.
- Numéricas: estandarización o *passthrough*.
- Target: balanceado (54 % sanos, 46 % enfermos).





- Modelo: regresión logística (liblinear).
- Métricas:
  - Accuracy: 0.90
  - Recall: 0.79
  - F1-score: 0.88
  - ROC-AUC: 0.96
- Matriz de confusión: 6 FN, 0 FP.



- Modelo robusto y explicable.
- Variables clave: *cp*, *thalach*, *oldpeak*, *ca*, *thal*.
- Mejoras futuras:
  - SMOTE para balanceo.
  - Ajustar umbral de decisión.
  - SHAP para interpretabilidad local.
- Recomendación: priorizar sensibilidad para entornos clínicos.



# sparkles ¡Gracias! heartbeat

*Predicción de Enfermedad Cardíaca con Ciencia de Datos*

