Clasificación de Riesgo Cardiovascular

Proyecto 2: Introducción a la Ciencia de Datos

Hazel Shamed Sánchez Chávez

Centro de Investigación en Matemáticas (CIMAT)

Contenido

- Introducción
- Exploración de datos
- Preprocesamiento
- Modelos de Clasificación
- Resultados y Comparación
- Conclusiones

Introducción

- Las enfermedades cardiovasculares (ECV) representan **32 % de todas las defunciones** globales (OMS)
- Identificación temprana: desafío crítico para sistemas de salud
- Dataset *"Heart Disease"* del repositorio UCI:
  - 303 pacientes
  - Registros clínicos y biomédicos
  - Signos clínicos, parámetros de laboratorio, características electrocardiográficas
- **Objetivo**: Clasificador binario para discriminar presencia/ausencia de enfermedad cardiovascular

Base de datos Heart Disease

- **Fuente**: Cleveland Clinic Foundation (Dr. Robert Detrano)
- **Período**: 1979-1988 (10 años)
- **Pacientes**: 303 referidos para evaluación coronaria
- **Criterio diagnóstico**: Angiografía coronaria ($\geq$50 % de estrechamiento)
- **Dimensión**: 303 filas $\times$ 13 columnas + etiqueta

Variables del Dataset

Variable	Tipo	Descripción
age	Continuo	Edad en años (29-77)
sex	Categórico	Sexo (0=fem, 1=masc)
cp	Categórico	Tipo de dolor torácico (1-4)
trestbps	Continuo	Presión arterial en reposo (mm Hg)
chol	Continuo	Colesterol sérico (mg/dl)
fbs	Categórico	Glicemia en ayunas >120 mg/dl
restecg	Categórico	Resultado ECG en reposo
thalach	Continuo	Frecuencia cardíaca máxima
exang	Categórico	Angina inducida por ejercicio
oldpeak	Continuo	Depresión ST inducida
slope	Categórico	Pendiente del segmento ST
ca	Continuo	Vasos principales coloreados
thal	Categórico	Tipo de defecto talio

Estadísticas Descriptivas

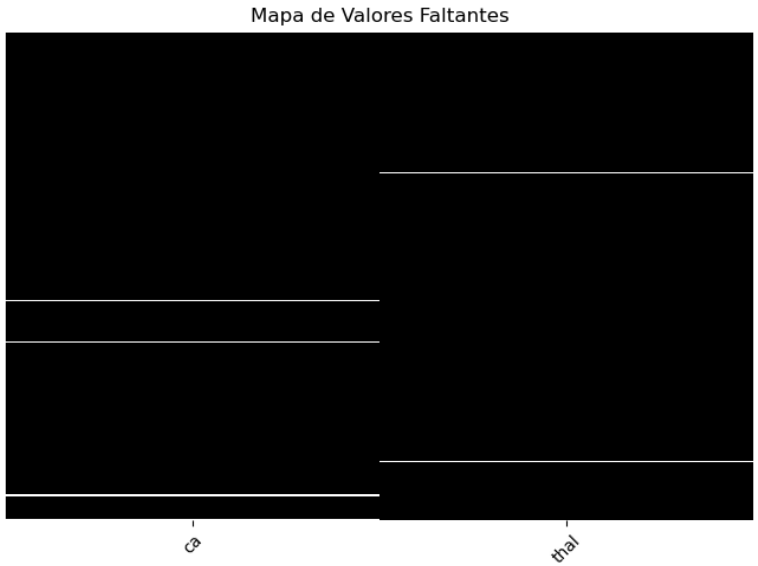
	age	trestbps	chol	thalach	oldpeak	ca
count	303	303	303	303	303	299
mean	54.4	131.69	246.69	149.61	1.04	0.67
std	9.04	17.60	51.78	22.88	1.16	0.94
min	29	94	126	71	0.00	0
max	77	200	564	202	6.20	3

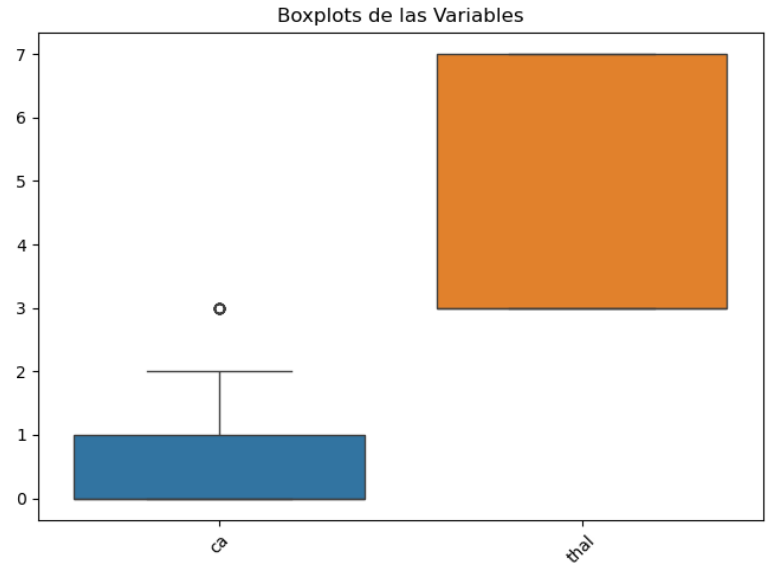
- **Distribución objetivo**: 164 sin enfermedad vs 139 con enfermedad
- **Valores faltantes**: en variables `ca` y `thal`
- **Consideraciones**: Datos históricos (1979-1988), sesgo de selección

Preprocesamiento de Datos

Objetivos

- Homogenizar escala de variables numéricas
- Separar predictores X y respuesta y
- Manejar valores faltantes y outliers

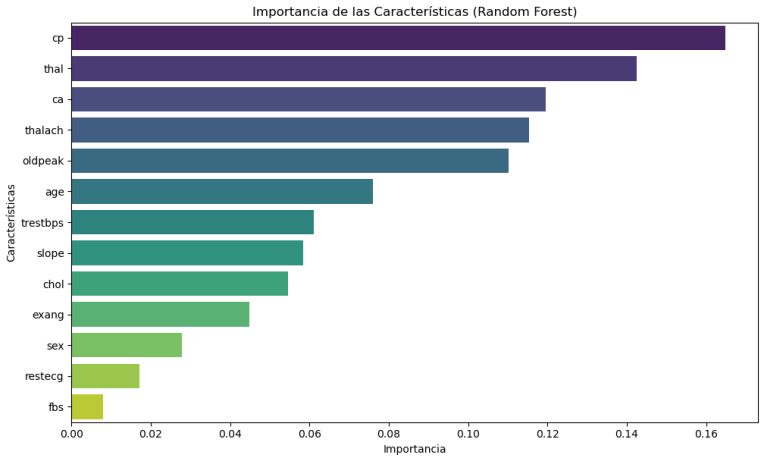




- **Valores faltantes**: Eliminación de filas (297 observaciones finales)
- **Escalado**: StandardScaler (media=0, varianza=1)

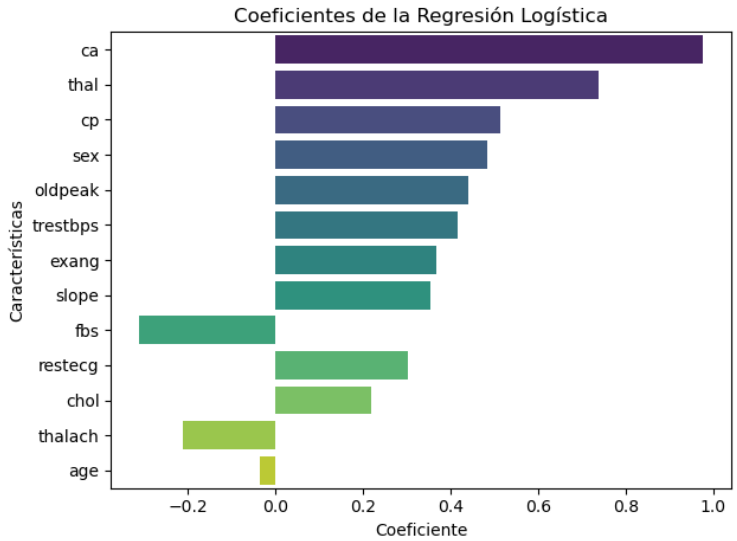
Random Forest

- Combina múltiples árboles de decisión
- Maneja variables categóricas/continuas
- Captura relaciones no lineales
- Proporciona importancia de variables



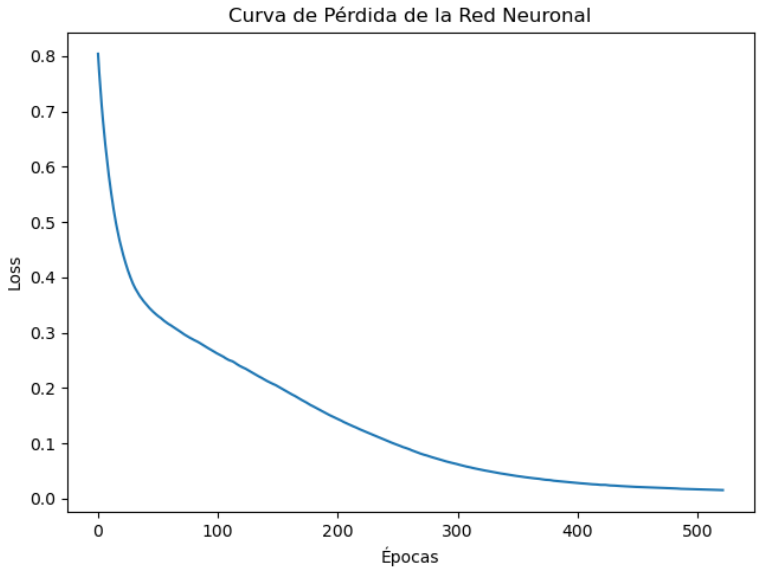
Regresión Logística

- Modelo lineal interpretable
- Coeficientes como *odds ratios*
- Bajo riesgo de sobreajuste
- Supone linealidad en el logit



Red Neuronal

- Aprende patrones no lineales complejos
- Flexible con arquitecturas específicas
- Requiere más datos y ajuste cuidadoso

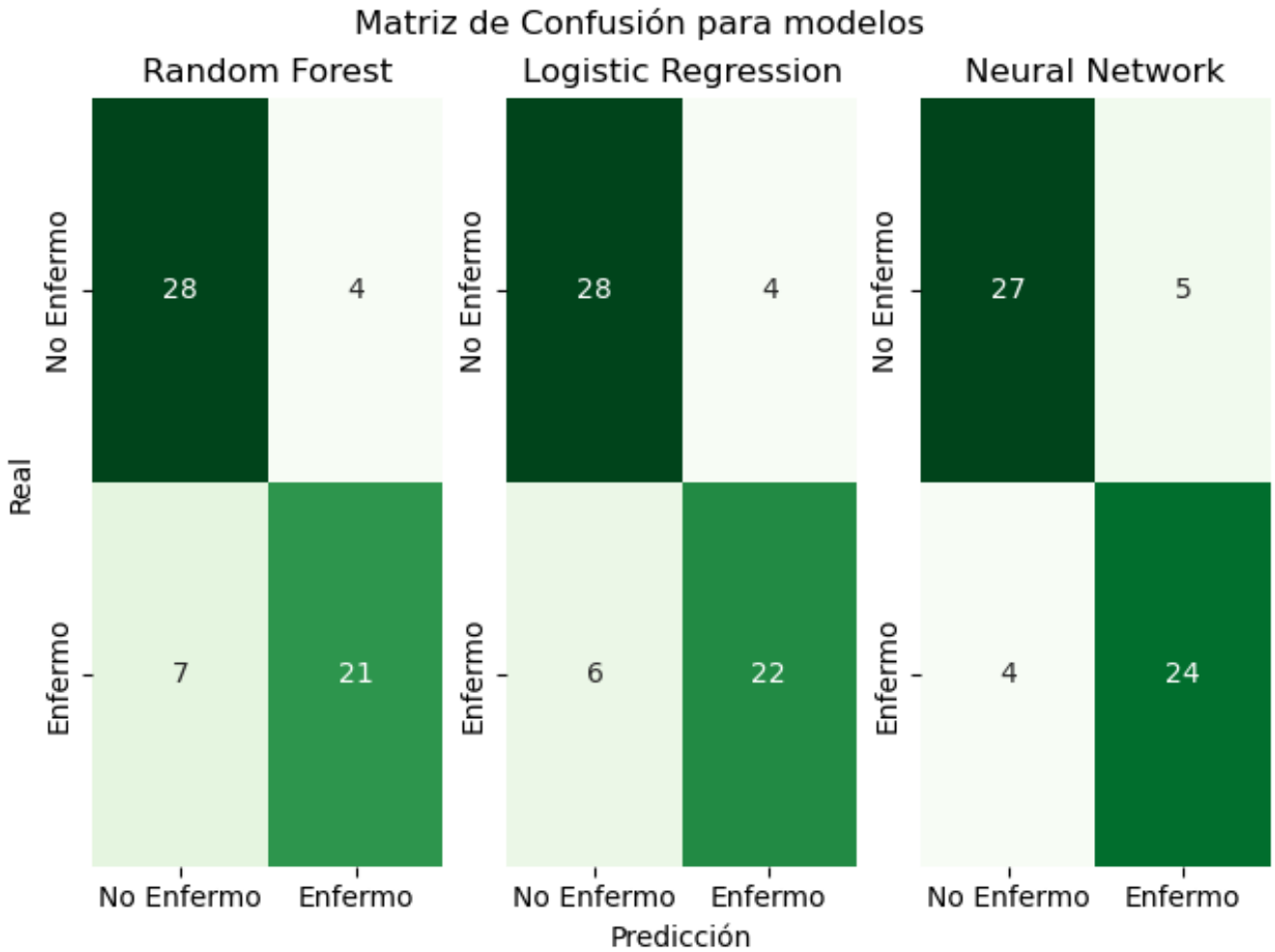


Métricas de Desempeño

Métrica	Random Forest	Logistic Regression	Neural Network
Recall	0.750	0.786	**0.857**
Accuracy	0.817	0.833	**0.850**
Balanced Accuracy	0.812	0.830	**0.850**
AUC-ROC	0.941	**0.950**	0.910
Precision	0.840	**0.846**	0.828
F1-score	0.792	0.815	**0.842**

- **Red Neuronal**: Mejor recall y accuracy (minimiza falsos negativos)
- **Regresión Logística**: Mejor AUC-ROC (capacidad discriminativa global)
- **Random Forest**: Posición intermedia en todas las métricas

Matrices de Confusión



- Red neuronal detecta mejor casos positivos (menos falsos negativos)
- Patrones de error diferentes entre modelos

Conclusiones

Hallazgos Principales

- **Regresión Logística**: Mejor capacidad discriminativa global (AUC-ROC 0.950)
- **Red Neuronal**: Mayor sensibilidad (recall 0.857) - ideal para minimizar falsos negativos
- **Random Forest**: Desempeño sólido pero intermedio

Recomendaciones Clínicas

- Prioridad en detección: Red Neuronal
- Equilibrio y interpretabilidad: Regresión Logística
- Alternativa robusta: Random Forest

Limitaciones y Trabajo Futuro

Limitaciones

- Datos históricos (1979-1988)
- Sesgo de selección (pacientes referidos)
- Valores faltantes en variables clave
- Desbalance de clases moderado

Trabajo Futuro

- Incorporar factores de riesgo modernos
- Validación en poblaciones contemporáneas
- Análisis de características por subpoblaciones
- Implementación en entorno clínico real

Referencias

- **Dataset**: Detrano, R. et al. (1988). *Heart Disease Dataset*. UCI Machine Learning Repository.
- **Organización Mundial de la Salud (2021)**. *Cardiovascular Diseases (CVDs)*.
- **Cleveland Clinic Foundation**. (1988). Multicenter cardiac study data.
- Hungarian Institute of Cardiology, University Hospital Zurich, University Hospital Basel. Contribuidores de datos.

¡Gracias por su atención!