Proyecto 2: Clasificación de Riesgo Cardiovascular
Introducción a la Ciencia de Datos

Estudiante: Hazel Shamed Sánchez Chávez
Institución: Centro de Investigación en Matemáticas (CIMAT)
Profesor: Dr. Marco Antonio Aquino López

1. Introducción

En el ámbito de la salud pública, las enfermedades cardiovasculares (ECV) constituyen una de las principales causas de mortalidad a nivel global, representando aproximadamente el 32 % de todas las defunciones según la Organización Mundial de la Salud. La identificación temprana de individuos en riesgo representa un desafío crítico para los sistemas de salud, con implicaciones significativas en la reducción de la carga asistencial y la mejora de resultados clínicos.

Este proyecto se enfoca en el desarrollo de un modelo predictivo de clasificación de riesgo cardiovascular utilizando el dataset *"Heart Disease"* del repositorio UCI Machine Learning, que contiene registros clínicos y biomédicos de 303 pacientes. La base de datos integra signos clínicos, parámetros de laboratorio y características electrocardiográficas, proporcionando una base multidimensional para el análisis.

Las ECV representan condiciones patológicas que afectan el corazón y los vasos sanguíneos, incluyendo enfermedad coronaria, enfermedad cerebrovascular y cardiopatía reumática. La naturaleza multifactorial de estas patologías demanda aproximaciones analíticas capaces de integrar diversos predictores, desde factores de riesgo convencionales hasta marcadores específicos de función cardíaca.

El objetivo central de este proyecto es construir un clasificador binario que permita discriminar entre pacientes con presencia y ausencia de enfermedad cardiovascular, utilizando únicamente variables accesibles mediante exámenes clínicos rutinarios. Este enfoque pretende superar las limitaciones de los métodos diagnósticos invasivos, optimizando la asignación de recursos diagnósticos especializados.

La implementación exitosa de este modelo podría servir como herramienta de apoyo a la decisión clínica, facilitando la estratificación prioritaria de pacientes y contribuyendo a estrategias de prevención secundaria más efectivas en el manejo de enfermedades cardiovasculares.

2. Exploración inicial de los datos

2.1. Base de datos *Heart Disease*.

El dataset "Heart Disease", de 303 filas por 13 columnas (mas etiquetas), fue recopilado principalmente por la **Cleveland Clinic Foundation** bajo la dirección del Dr. Robert Detrano, M.D., Ph.D., con la colaboración de múltiples instituciones internacionales incluyendo el Hungarian Institute of Cardiology de Budapest, University Hospital de Zurich y University Hospital de Basel. Los datos fueron donados al UCI Machine Learning Repository en 1988, donde están disponibles públicamente con el identificador 45.

El estudio se condujo como una investigación observacional multicéntrica entre 1979 y 1988, abarcando aproximadamente 10 años de recolección de datos. La población del estudio consistió en 303 pacientes referidos para evaluación coronaria en las instituciones participantes. El criterio de referencia para el diagnóstico fue la angiografía coronaria, definiendo como caso positivo la presencia de $\geq$50 % de estrechamiento diametral.

El dataset original contenía 76 variables recolectadas, pero para el análisis estándar se seleccionaron 14 atributos que incluyen signos clínicos, parámetros de laboratorio y características electrocardiográficas.

2.2. Variables del Dataset

Al descargar el dataset, las variables categoricas ya vienen dadas de forma numerica, asignando un valor entero a cada categoría de cada variable. Esto facilita la manipulación y calculos de los datos. Presentamos una tabla resumen de las variables del dataset:

Cuadro 1: Variables del Dataset Heart Disease

Atributo	**Tipo**	**Descripción**	**Valores**
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)	94-200
chol	Continuo	Colesterol sérico (mg/dl)	126-564
fbs	Categórico	Glicemia en ayunas >120 mg/dl	0 = no, 1 = sí
restecg	Categórico	Resultado ECG en reposo	0-2
thalach	Continuo	Frecuencia cardíaca máxima alcanzada	71-202
exang	Categórico	Angina inducida por ejercicio	0 = no, 1 = sí
oldpeak	Continuo	Depresión ST inducida por ejercicio	0-6.2
slope	Categórico	Pendiente del segmento ST en ejercicio	1-3
ca	Continuo	Número de vasos principales coloreados	0-3
thal	Categórico	Tipo de defecto talio	3 = normal, 6 = defecto, 7 = reversible
num	Objetivo	Diagnóstico de enfermedad	0 = no, 1-4 = sí

Estos signos clínicos, parámetros de laboratorio y características electrocardiográficas se interpretan desde el punto de vista médico de la siguiente manera:

- El dolor torácico (cp) se clasifica en cuatro tipos: anginoso típico (dolor opresivo por esfuerzo), atípico, no anginoso y ausente.
- El colesterol sérico (chol) cuantifica los lípidos en sangre en mg/dl.
- La glicemia en ayunas (fbs) indica diabetes si supera $120 mg/dl$.
- El electrocardiograma en reposo (restecg) detecta anomalías ST-T o hipertrofia ventricular.
- La frecuencia cardíaca máxima (thalach) registra el pico de esfuerzo.
- La angina inducida por ejercicio (exang) señala dolor durante la prueba de esfuerzo.
- La depresión del segmento ST (oldpeak) mide en milivolts la isquemia miocárdica durante el ejercicio.
- La pendiente del segmento ST (slope) evalúa si es ascendente (normal), plana o descendente (patológica).
- Los vasos principales coloreados (ca) indican cuántas arterias coronarias tienen obstrucción 50 % en angiografía.
- La prueba de talio (thal) evalúa perfusión cardíaca: normal, defecto fijo (infarto) o reversible (isquemia).
- El diagnóstico (num) clasifica la severidad de la enfermedad desde 0 (ausente) hasta 4 (severa).

A continuación, se presentan las estadísticas descriptivas de estas variables:

Cuadro 2: Estadísticas Descriptivas del Dataset

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
count	303	303	303	303	303	303	303	303	303	303	303	299	301
mean	54.4	0.68	3.16	131.69	246.69	0.15	0.99	149.61	0.33	1.04	1.60	0.67	4.73
std	9.04	0.47	0.96	17.60	51.78	0.36	0.99	22.88	0.47	1.16	0.62	0.94	1.94
min	29	0	1	94	126	0	0	71	0	0.00	1	0	3
25 %	48	0	3	120	211	0	0	134	0	0.00	1	0	3
50 %	56	1	3	130	241	0	1	153	0	0.80	2	0	3
75 %	61	1	4	140	275	0	2	166	1	1.60	2	1	7
max	77	1	4	200	564	1	2	202	1	6.20	3	3	7

2.3. Consideraciones Adicionales

Es importante señalar que el dataset presenta algunos valores faltantes en los atributos `ca` y `thal`. La distribución de la variable objetivo muestra que de los 303 pacientes, 164 no presentaban enfermedad cardiovascular (valor 0), mientras que 139 presentaban diversos grados de enfermedad (valores 1-4), esto indica que las clases son desbalanceadas.

Entre las limitaciones del dataset se debe considerar su contexto temporal (datos de 1979-1988), donde los criterios diagnósticos y prácticas clínicas pueden diferir de los estándares actuales. Además, la muestra representa una población específica de pacientes referidos para angiografía, lo que puede introducir sesgos de selección. Cabe resaltar que el dataset no incluye factores de riesgo modernos como proteína C reactiva o score cálcico coronario.

3. Preprocesamiento

El preprocesamiento tiene como finalidad transformar la base de datos *Heart Disease* en una matriz de diseño adecuado para el modelado supervisado. En particular, buscamos,

a) homogenizar la escala de las variables numéricas para algoritmos sensibles a la magnitud,

b) separar correctamente predictores X y respuesta y evitando fugas de información.

3.1. Preparación y limpieza

Como mencionamos anteriormente, las variables `ca` y `thal` presentan valores faltantes. A continuación, mostramos de forma grafica como se distribiuyen estos valores faltantes en el dataset y a su vez analizamos la posible presencia de outliers a través de boxplots.

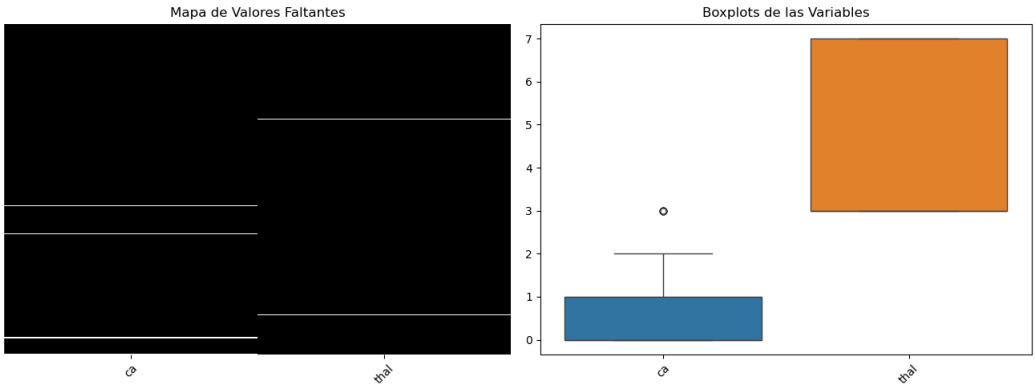

Figura 1: Valores faltantes y boxplots de `ca` y `thal`

A simple vista observamos que los datos faltes no siguen un patrón específico, por lo que optamos por eliminar las filas con valores faltantes, resultando en un total de 297 observaciones. En cuanto a los outliers, considerando tambien la siguiente tabla, no se observan valores extremos evidentes que justifiquen una eliminación, por lo que los conservamos para el analisis posterior.

Cuadro 3: Distribución de Variables Categóricas

Variable	**Valor**	**Frecuencia**
ca	0.0	176
	1.0	65
	2.0	38
	3.0	20
thal	3.0	166
	7.0	117
	6.0	18

Por otro lado, se seleccionó StandardScaler para el preprocesamiento de los datos debido a su idoneidad en el contexto de variables biomédicas. Este método transforma las características restándoles la media y escalándolas por la desviación estándar, resultando en una distribución con media cero y varianza unitaria. Esta estandarización es crucial para algoritmos sensibles a la escala, como la regresión logística y las redes neuronales, ya que acelera la convergencia y asegura que los coeficientes o pesos se ajusten de manera equilibrada. A diferencia de MinMaxScaler, que es sensible a valores atípicos, StandardScaler preserva la forma original de la distribución y maneja mejor los rangos variables de las características (como colesterol: 126-564 y presión arterial: 94-200). Dado que los valores atípicos en datos biomédicos suelen ser clínicamente relevantes y no errores de medición, StandardScaler resulta el balance óptimo entre robustez y preservación de la información.

4. Clasificación y evaluación de modelos

Basados en las propiedades metodológicas y las ventajas prácticas que cada clasificador ofrece para el análisis de riesgo de enfermedades cardiovasculares, hemos elejido implementar tres clasificadores. Estos son: bosques aleatorios, regresión logística y redes neuronales.

4.1. Random Forest

El modelo *Random Forest* combina múltiples árboles de decisión para mejorar la precisión y estabilidad del clasificador. Es adecuado para datos clínicos porque maneja variables categóricas y continuas sin preprocesamientos complejos, captura relaciones no lineales y ofrece medidas de importancia de variables que apoyan la interpretación.

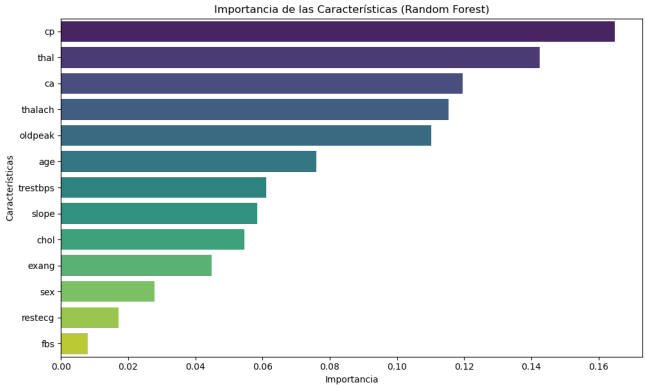

Figura 2: Importancia de características en Random Forest

4.2. Regresión Logística

La regresión logística es un modelo lineal ampliamente utilizado en clasificación binaria por su sencillez e interpretabilidad. Proporciona coeficientes en términos de *odds ratios*, genera probabilidades predichas adecuadas para definir umbrales clínicos y presenta bajo riesgo de sobreajuste, lo que la convierte en una línea base robusta. Su principal limitación es la suposición de linealidad en el logit.

La importancia de las variables según los coeficientes del modelo se muestra en la siguiente figura:

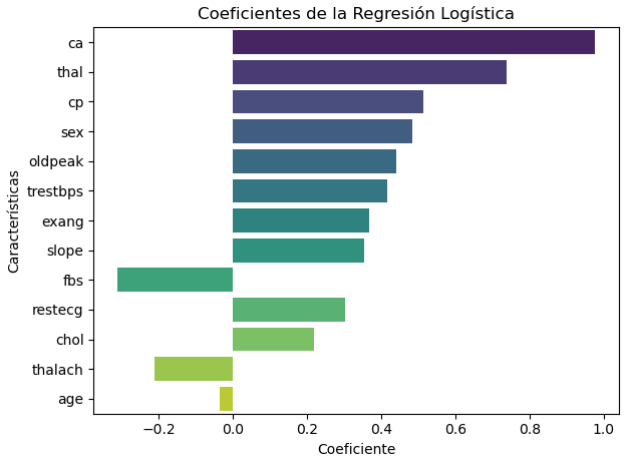

Figura 3: Coeficientes de la Regresión Logística

4.3. Red Neuronal

Las redes neuronales son modelos flexibles capaces de aprender patrones no lineales complejos y adaptarse mediante arquitecturas específicas y técnicas de regularización como *dropout*. Resultan útiles cuando existen interacciones difíciles de capturar con modelos tradicionales. Sin embargo, requieren más datos y cuidado en su ajuste debido al riesgo de sobreajuste y su menor interpretabilidad.

Durante el entrenamiento, se monitoreó la función de pérdida para evaluar la convergencia del modelo, como se observa en la siguiente figura:

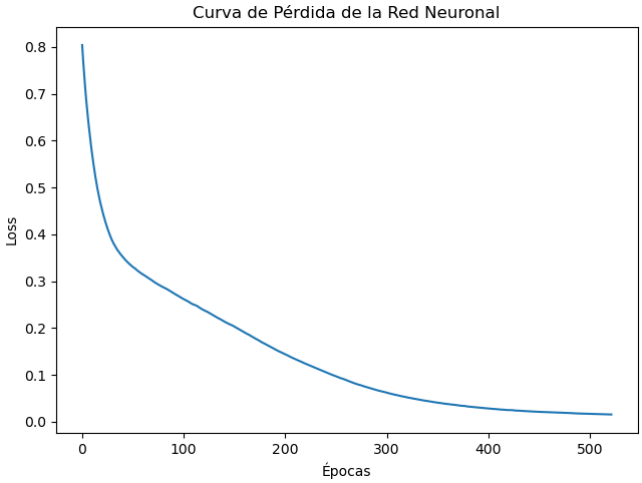

Figura 4: Curva de pérdida durante el entrenamiento de la Red Neuronal

4.4. Comparación de los modelos

Las gráficas de importancia de características en el bosque aleatorio evidencia qué variables aportan mayor poder predictivo. Por otro lado, los coeficientes de la regresión logística permiten entender la dirección e influencia de cada variable en el riesgo estimado. La curva de pérdida de la red neuronal confirma un proceso de aprendizaje estable sin señales evidentes de sobreajuste.

Mostramos las siguientes mmétricas que se obtuvieron en cada modelo:

Cuadro 4: Métricas de desempeño para los clasificadores

Métrica	**Random Forest**	**Logistic Regression**	**Neural Network**
Recall (Sensitivity)	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

Los tres modelos muestran un desempeño sólido, pero con diferencias relevantes según la métrica. La red neuronal alcanza la mayor recall (0.857) y la mayor accuracy y balanced accuracy (0.850), lo que indica que es la opción más efectiva para detectar casos positivos (minimizar falsos negativos), aspecto crítico en un problema clínico donde no detectar a un paciente enfermo es costoso. Sin embargo, su AUC-ROC (0.910) es la más baja de los tres, lo que sugiere que, globalmente, discrimina algo peor entre clases a través de todos los umbrales que la regresión logística (AUC 0.950). La regresión logística presenta el mejor AUC-ROC y buena precisión (0.846) y F1 (0.815), lo que la convierte en un modelo muy equilibrado y con excelente capacidad de discriminación (útil si se busca un compromiso entre detectar enfermos y limitar falsos positivos). El bosque aleatorio queda en posición intermedia: buena AUC (0.941) y precisión (0.840), pero menor recall (0.750), por lo que produce más falsos negativos que la red neuronal.

Por lo tanto, si la prioridad clínica es maximizar la detección (minimizar FN) es conveniente usar la red neuronal; si se prioriza la capacidad de discriminación global y estabilidad interpretativa, la regresión logística es la opción preferible; el bosque aleatorio es una alternativa intermedia.

A continuación, se presentan las matrices de confusión para cada modelo:

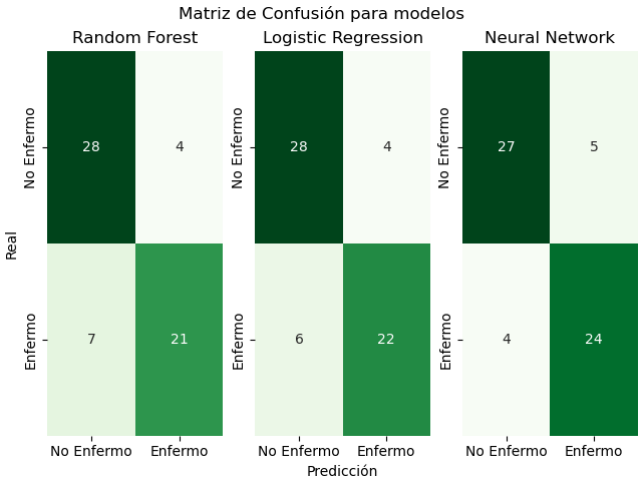

Figura 5: Matrices de confusión para los clasificadores

Las matrices de confusión ilustran claramente la diferencia en patrones de error entre modelos, destacando el mejor desempeño de la red neuronal en la detección de casos positivos.

5. Conclusiones

El análisis desarrollado sobre el conjunto de datos clínicos permitió evaluar el desempeño de tres modelos de clasificación orientados a la detección de riesgo cardiovascular, identificando fortalezas, limitaciones y su utilidad potencial en un contexto aplicado. Los resultados muestran que la regresión logística ofrece la mayor capacidad discriminativa global con una AUC-ROC cercana a 0,95, lo que la convierte en el modelo más estable cuando se requiere equilibrio entre sensibilidad y especificidad. Por su parte, la red neuronal presenta la sensibilidad más alta del estudio y alcanza la mejor exactitud y balanced accuracy, lo cual la posiciona como la alternativa más adecuada cuando la prioridad clínica es reducir al mínimo los falsos negativos y asegurar que la mayoría de los pacientes con riesgo sean identificados oportunamente. El modelo de bosques aleatorio, aunque competitivo, tiende a generar un mayor número de falsos negativos respecto a la red neuronal, situándose en un punto intermedio tanto en desempeño como en robustez.

No obstante, el estudio presenta limitaciones relevantes: el conjunto de datos es antiguo y no necesariamente representativo de poblaciones actuales, existe presencia de valores faltantes y cierto desbalance de clases, y los datos provienen de pacientes referidos para procedimientos cardiovasculares, lo que puede introducir sesgos de selección.

Referencias

- **Dataset:** Detrano, R. et al. (1988). *Heart Disease Dataset*. UCI Machine Learning Repository. Recuperado el 25 de noviembre de 2025 de https://archive.ics.uci.edu/ml/datasets/Heart+Disease
- **Organización Mundial de la Salud (2021).** *Cardiovascular Diseases (CVDs)*. Recuperado de https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Nota: Las ECV representan el 32 % de las muertes globales