Parte 1 del laboratorio de automatización.

1. **. Análisis preliminar del problema.**

a. Determine si se trata de un **problema de clasificación o regresión**. Justifique su respuesta e indique claramente el target (variable objetivo).

Se trata de un problema de clasificación binaria, en donde el objetivo es realizar un modelo que sea capaz de predecir riesgo coronario a 10 años. El problema sugiere un modelo de clasificación porque queremos predecir una categoría con dos posibles valores (0 y 1) estos valores al ser enteros sugieren un problema de clasificación.

Justificación: El modelo no busca estimar un valor numérico continuo, sino una probabilidad de que ocurra un evento como el riesgo coronario, el target es la columna TenYearCHD .

b. Clasifique las características en tipos de variables (numéricas, categóricas, binarias, ordinales, etc.).

Tipo	**Variables**	**Descripción**
Binarias	male, currentSmoker, BPMeds, prevalentStroke, prevalentHyp, diabetes, TenYearCHD	Variables que solo toman valores 0/1 (sí/no, presencia/ausencia).
Discretas	education, cigsPerDay	Valores enteros contables (años de educación, número de cigarrillos).
Continuas	age, totChol, sysBP, diaBP, BMI, heartRate, glucose	Variables medidas en una escala numérica continua (presión, colesterol, glucosa, etc.).

c. Investigue y explique el **protocolo de adquisición y/o generación de datos** que siguieron los investigadores.

El dataset proviene del Framingham Heart Study, un estudio cardiovascular en curso con residentes de Framingham, Massachusetts, cuyo objetivo es predecir el riesgo de desarrollar enfermedad coronaria a 10 años.

Para la adquisición de datos, los investigadores realizaron una selección de participantes dentro de un grupo demográfico de 30 a 70 años. Posteriormente, se aplicaron diversos exámenes médicos y clínicos con el fin de registrar variables estrechamente relacionadas con el riesgo coronario, tales como presión arterial, niveles de colesterol, glucosa, tabaquismo, entre otros factores.

Entrenamiento de modelos

Modelo	F1 (train)	F1 (val)	F1 (test)	Accuracy (train)	Accuracy (val)	Accuracy (test)
KNN	0.1514	0.1101	0.0594	0.8564	0.8475	0.8506
DNN	0.0	0.0	0.0	0.8479	0.8475	0.8491
Random Forest	0.7195	0.246	0.1522	0.9177	0.7783	0.7547

Estrategias de Validación de Modelos

1. K-Fold Cross Validation.

K-Fold Cross Validation divide el dataset en K subconjuntos (folds) de tamaño similar. El modelo se entrena K veces, usando K-1 folds para entrenamiento y 1 fold para validación en cada iteración. Al final, se promedian las métricas para obtener un resultado más robusto.

2. Leave-One-Out Cross Validation (LOOCV)

LOOCV es una variante extrema de K-Fold, donde K = N (el número total de observaciones). En cada iteración, el modelo se entrena con N-1 observaciones y se valida con 1 sola observación. Se repite para todas las observaciones del dataset.

3. Análisis.

para el análisis de riesgo coronario, **k-fold cross validation es la estrategia más recomendable**, ya que combina eficiencia y robustez en la evaluación, mientras que LOOCV, aunque teóricamente válido, resulta poco práctico y costoso computacionalmente. El uso de estas técnicas proporciona una **evaluación más confiable del modelo** y permite tomar decisiones más informadas sobre su aplicabilidad en escenarios reales de predicción de riesgo cardiovascular.

Diagramsa de flujo del pipeline.