Estadística para las biociencias

Gonzalo Jaén, Pau Nerín, Jovan Pomar y Juan Ignacio Sampere
(JANEPOSA)

17 de junio de 2025

Introducción

- La **diabetes mellitus tipo 2 (DM2)** es una enfermedad crónica caracterizada por *hiperglucemia persistente*.
- Afecta a millones de personas en todo el mundo y representa un importante problema de salud pública.
- Su detección temprana permite intervenir con cambios en el estilo de vida o tratamiento farmacológico, reduciendo el riesgo de complicaciones.

Contexto clínico y dataset

En este estudio se emplea el **Pima Indians Diabetes Dataset**, que contiene datos de 768 mujeres pima mayores de 21 años, incluyendo:

- Datos fisiológicos y resultados médicos rutinarios.
- Variables numéricas sin necesidad de codificación categórica.
- Variable objetivo: **Outcome** (0 = No diabética, 1 = Diabética).

Variables del estudio

Variable	**Descripción**
Pregnancies	Número de embarazos
Glucose	Glucosa en ayunas (mg/dL)
BloodPressure	Presión diastólica (mm Hg)
SkinThickness	Grosor del tríceps (mm)
Insulin	Insulina sérica (μU/mL)
BMI	Indice de masa corporal (kg/m^2)
DiabetesPedigreeFunction	Riesgo hereditario
Age	Edad (años)
Outcome	Diagnóstico DM2 (1 = Sí, 0 = No)

Objetivos del estudio

- Evaluar la capacidad de clasificación de los modelos **k-Nearest Neighbours (k-NN)** y **Support Vector Machines (SVM)**.
- Comparar su rendimiento en base a:
  - Matriz de confusión: sensibilidad, especificidad, accuracy, kappa.
  - Curva ROC y **Area Under the Curve (AUC)**.
- Determinar el modelo más apropiado para cribado clínico inicial.

Objetivos del estudio

- Evaluar la capacidad de clasificación de los modelos **k-Nearest Neighbours (k-NN)** y **Support Vector Machines (SVM)**.
- Comparar su rendimiento en base a:
  - Matriz de confusión: sensibilidad, especificidad, accuracy, kappa.
  - Curva ROC y **Area Under the Curve (AUC)**.
- Determinar el modelo más apropiado para cribado clínico inicial.

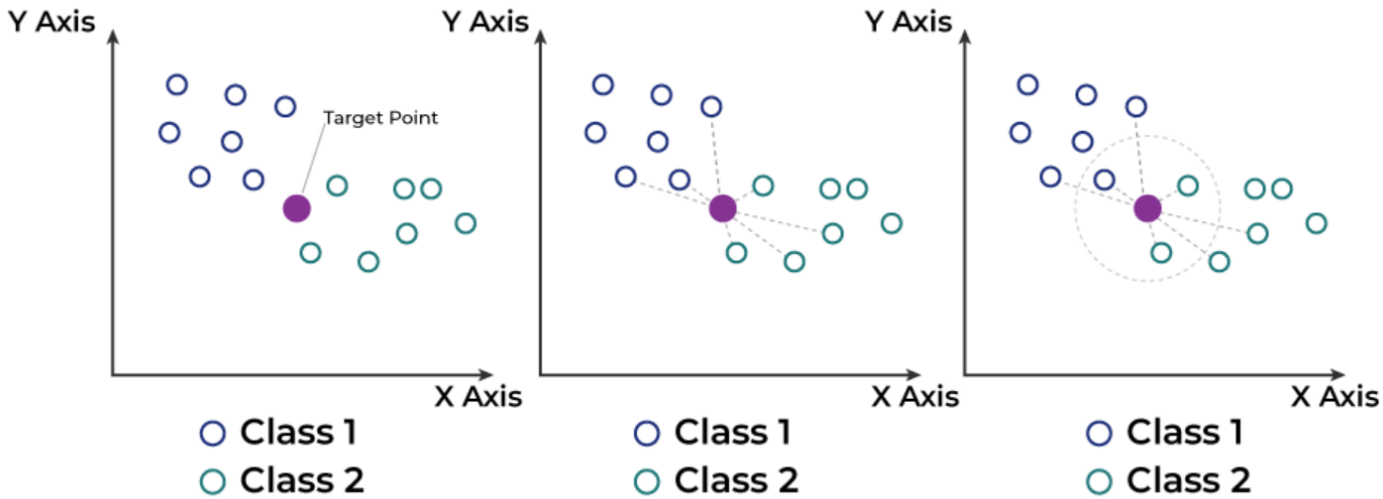
Ambos algoritmos de **aprendizaje supervisado** intentan aproximar la función de decisión:

$$f : \mathbb{R}^p \to \{0, 1\}, \qquad \hat{y} = f(\mathbf{x}),$$

donde $\mathbf{x} = (x_1, \ldots, x_p)$ representa el vector de predictores clínicos y $\hat{y}$ el diagnóstico estimado ($1 =$ diabética, $0 =$ no diabética).

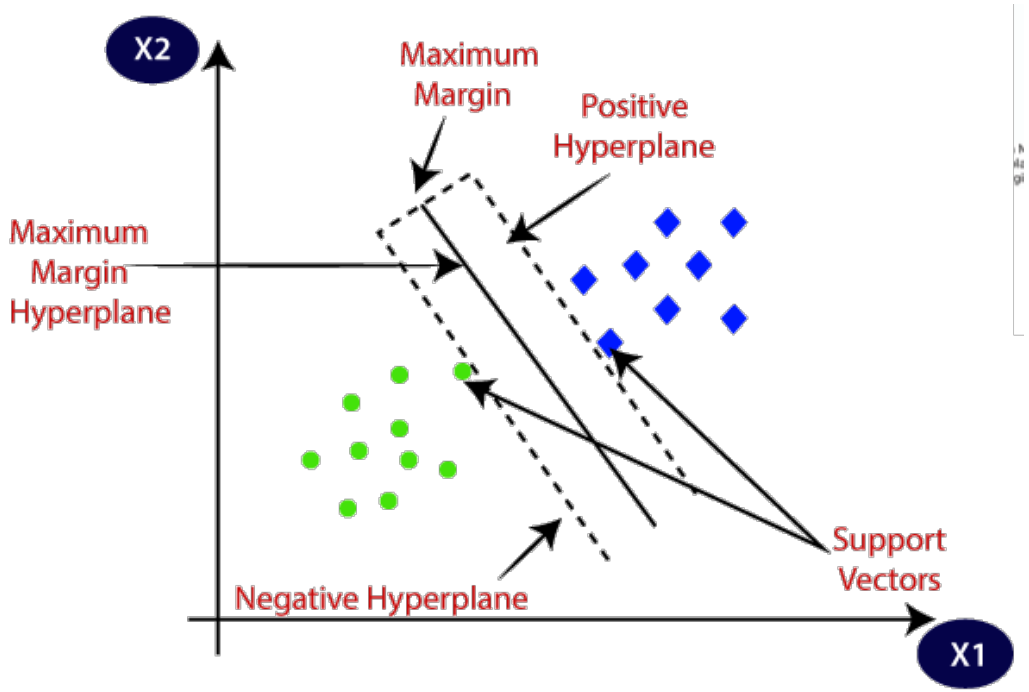
k-NN: Principio básico

- Clasifica una nueva observación según la clase más común entre sus **k vecinos más cercanos**.
- Utiliza distancia euclídea sobre datos estandarizados.
- Permite fronteras no lineales, sensibles al valor de k.

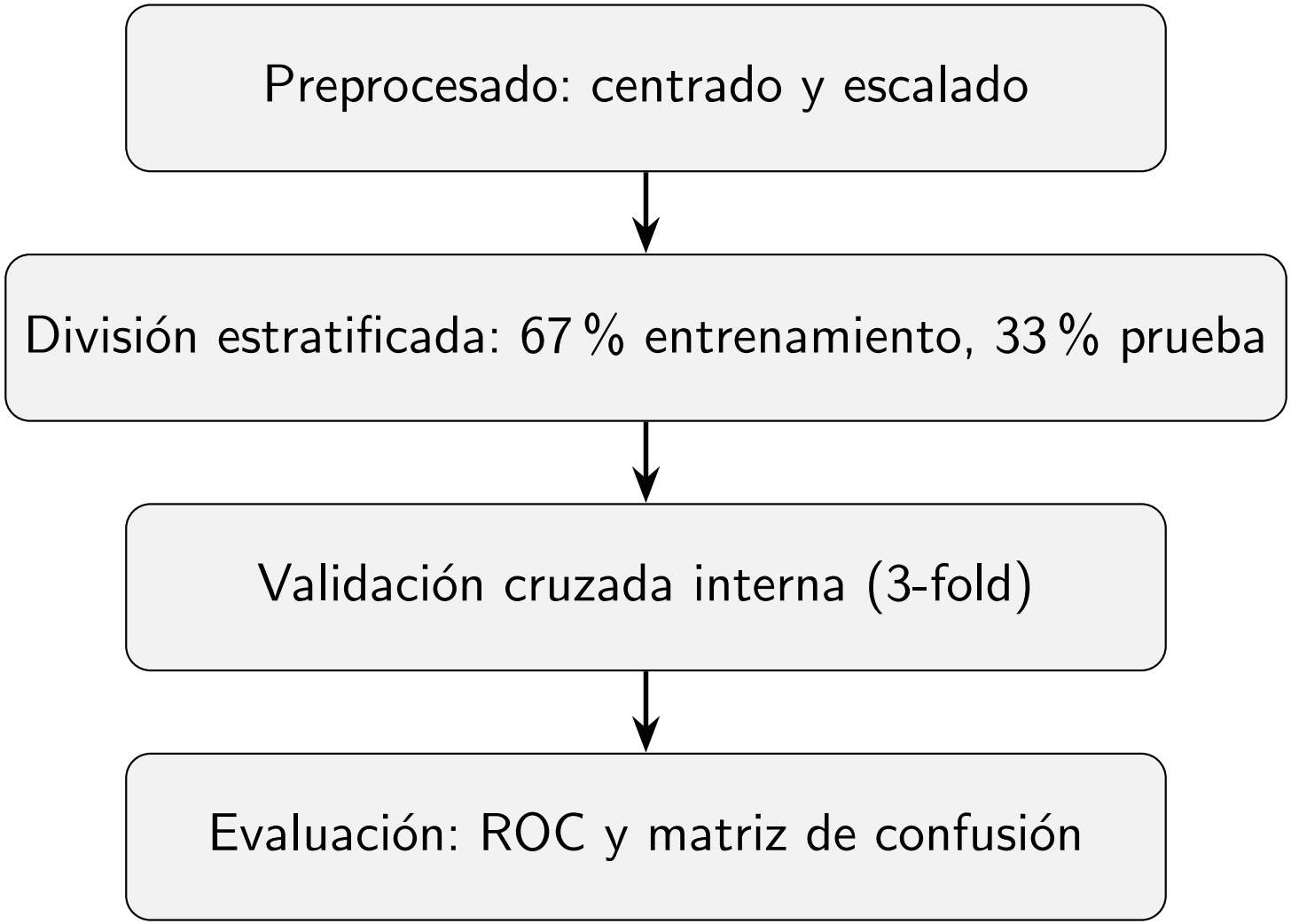


SVM: Principio básico

- Encuentra el hiperplano que maximiza la separación entre clases.
- Puede ser **lineal** o usar **kernels no lineales** como RBF.
- Introduce parámetros como C (penalización) y γ (flexibilidad del kernel).



Metodología



Resumen de Resultados del Modelo k-NN

Matriz de Confusión

Referencia	No	Sí
Predicción No	153	48
Predicción Sí	12	40

Estadísticos clave

Métrica	Valor
Accuracy	0.7628
Kappa	0.4221
McNemar p-valor	$6{,}228 \times 10^{-6}$

Métricas de rendimiento

Métrica	Valor
Sensibilidad (Recall)	0.4545
Especificidad	0.9273
Valor predictivo positivo (PPV)	0.7692
Valor predictivo negativo (NPV)	0.7612
Prevalencia	0.3478
Tasa de detección	0.1581
Prevalencia predicha	0.2055
Precisión balanceada	0.6909

Resumen de Resultados del Modelo SVM

Matriz de Confusión

Referencia	No	Sí
Predicción No	140	37
Predicción Sí	25	51

Estadísticos clave

Métrica	Valor
Accuracy	0.7549
Kappa	0.4421
McNemar p-valor	0.1624

Métricas de rendimiento

Métrica	Valor
Sensibilidad (Recall)	0.5795
Especificidad	0.8485
Valor predictivo positivo (PPV)	0.6711
Valor predictivo negativo (NPV)	0.7910
Prevalencia	0.3478
Tasa de detección	0.2016
Prevalencia predicha	0.3004
Precisión balanceada	0.7140

Comparación de Modelos de Clasificación

Resumen comparativo entre SVM-Lineal y k-NN ($k = 31$):

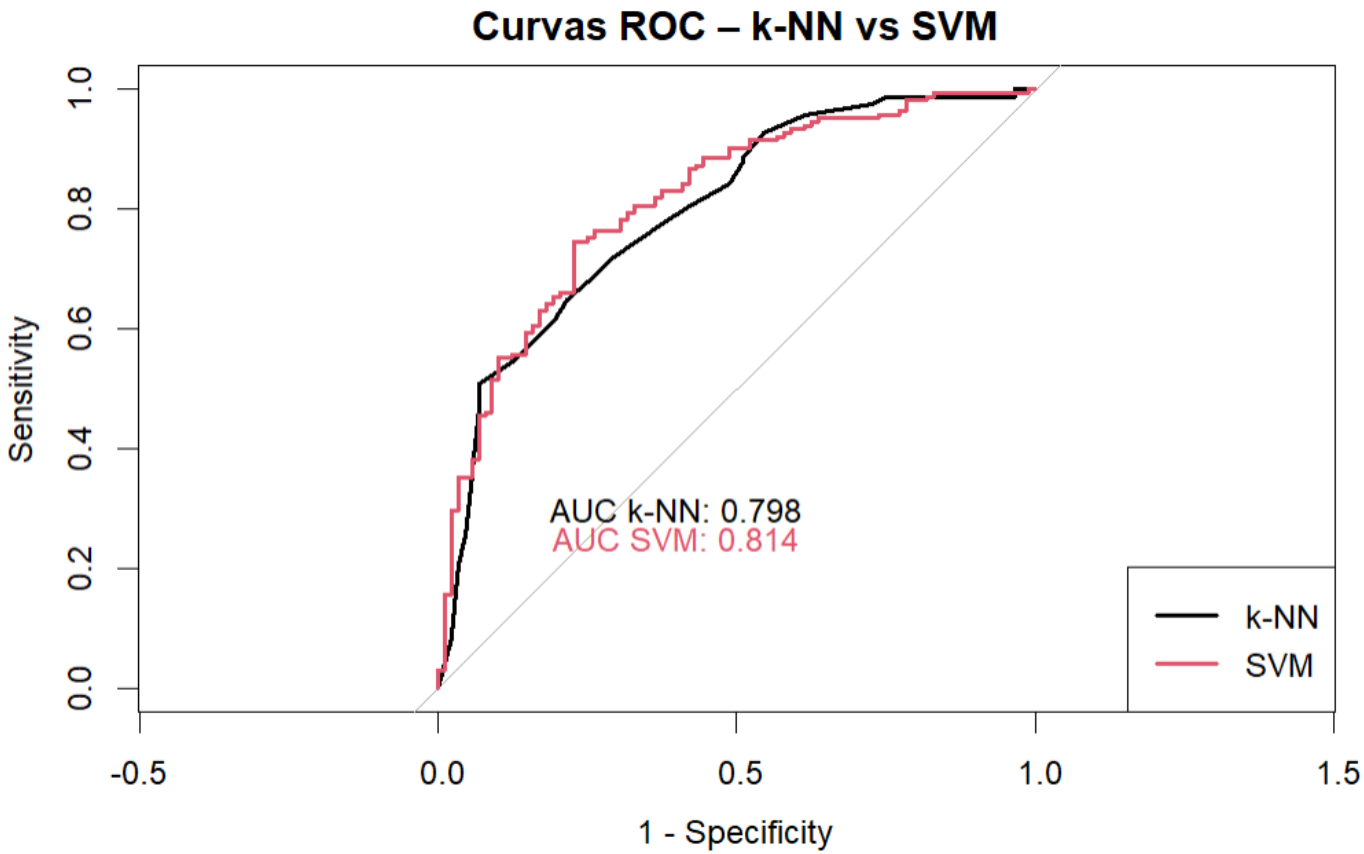
Modelo	Accuracy	Kappa	Sensibilidad	Especificidad	AUC
SVM-Lineal	0.755	0.442	0.580	0.848	0.814
k-NN ($k = 31$)	0.763	0.422	0.455	0.927	0.798

- **SVM-Lineal** tiene mejor **AUC** y **sensibilidad**, lo que lo hace más sensible a positivos.
- **k-NN ($k = 31$)** ofrece mayor **accuracy** y **especificidad**, ideal para evitar falsos positivos.

Modelo	TN	FP	FN	TP
k-NN ($k = 31$)	153	**12**	48	40
SVM-Lineal	140	25	**37**	**51**

Curvas ROC

- La curva ROC de SVM se mantiene por encima de la de k-NN.
- Ventaja de SVM en detección temprana.



Discusión: Síntesis de resultados y limitaciones

Síntesis de resultados:

- Ambos modelos alcanzan **accuracy** 0.76.
- **SVM-Lineal**: mayor **sensibilidad (0.58)** y **AUC (0.814)** mejor detección de casos.
- **k-NN (k=31)**: mayor **especificidad (0.93)** menos falsos positivos.
- Curva ROC: SVM domina en la mayoría de umbrales; k-NN sobresale en FPR ¡ 0.1.

Limitaciones:

1. **Representatividad:** población solo femenina y pima.
2. **Tamaño muestral:** limitado para generalizar.
3. **Variables:** faltan indicadores clínicos relevantes (HbA1c, fármacos).
4. **Validación:** solo una partición externa, 3-fold interna.
5. **Umbral fijo (0.5):** sin calibración basada en coste/prevalencia.

Discusión: Elección del modelo y uso clínico

Criterios de decisión clínica

- Priorizar **alta sensibilidad** para evitar falsos negativos (FN) balanceando especificidad.
- Considerar el **costo clínico** y **operativo** de errores tipo I y II.

Elección operativa

SVM-Lineal es preferible cuando:

- Se busca **detectar todos los casos posibles** (cribado inicial).
- El coste de omitir un paciente con diabetes es elevado.

Alternativa conservadora

k-NN (k = 31) es útil cuando:

- Se desea evitar falsos positivos.
- Los recursos son limitados o el diagnóstico es invasivo/caro.

Conclusiones

- Ambos modelos tienen puntos fuertes.
- SVM ideal para primer cribado poblacional.
- k-NN valioso para evitar derivaciones innecesarias.

Propuesta híbrida

- 1^{er} paso: **SVM-Lineal** para detección amplia (alta sensibilidad).
- 2^{do} paso: **k-NN** para refinar positivos (alta especificidad).
- Posibilidad de calibración de umbrales o modelos combinados (ensembles).

Referencias bibliográficas

1. Zar, J. H. (2010). *Biostatistical Analysis*. Pearson Education.

2. Smith, A., & Johnson, B. (2015). Advances in Bioinformatics: Genomic Data Analysis. *Bioinformatics Journal*, **31**(12), 1987–1995.

3. Muller, S., & Lee, D. (2020). *Machine Learning for Biomedical Applications*. Springer.