Resumen de comparación de los modelos

Se entrenaron y evaluaron tres modelos de clasificación: Random Forest Basico, Random Forest con reducción de complejidad, XGBoost y una variante de XGBoost con el parámetro `scale\_pos\_weight` ajustado para mejorar el rendimiento frente al desbalance de clases.

# 1. Random Forest (Basico) Este modelo muestra una señal clara de sobreajuste (overfitting). Presenta un rendimiento perfecto en el conjunto de entrenamiento, clasificando correctamente todos los casos, con precisión, recall y f1-score de 1.00 para ambas clases. Sin embargo, al evaluarlo con datos nuevos, el rendimiento disminuye considerablemente: el accuracy baja al 79.2% y el recall para la clase 1 (fumadores) cae a 0.71. Esto significa que solo logra detectar al 71% de los fumadores, lo cual no está mal, pero está lejos del 100% que obtenía en entrenamiento. Este comportamiento indica que el modelo aprendió demasiado bien los datos originales, sin lograr generalizar adecuadamente a nuevos casos.

# 2. Random Forest (Complejidad reducida)

En el conjunto de entrenamiento alcanzó un accuracy del 76%, con buen recall para la clase 1 (0.94), pero un rendimiento más bajo para la clase 0.   
Esto significa que el modelo tiene buena capacidad para identificar correctamente a los fumadores, pero también tiende a confundir a muchos no fumadores como si lo fueran (falsos positivos).  
En el conjunto de test, el accuracy bajó al 73%, manteniendo un recall alto (0.91) para la clase 1, pero con una menor precisión general (f1-score global ≈ 0.74).

# 3. XGBoost (modelo estándar)

Obtuvo los mejores resultados generales. En entrenamiento logró un accuracy del 86%, con buen balance entre precisión y recall para ambas clases.  
En test mantuvo una buena performance (accuracy de 77%), superando a Random Forest tanto en precisión como en recall para ambas clases.  
El modelo no solo logró buenos resultados en los datos con los que se entrenó, sino que también mantuvo un rendimiento consistente y estable cuando se lo evaluó con datos nuevos que no había visto antes. Esto indica que tiene una buena capacidad de generalización y es menos propenso a sobreajustar.

# 4. XGBoost con scale\_pos\_weight

Este modelo fue ajustado para favorecer el reconocimiento de la clase menos representada (los fumadores).  
En entrenamiento bajó un poco el accuracy (83%), pero mejoró el recall de la clase 1 (0.93), es decir, detectó a más fumadores correctamente.  
En test obtuvo un accuracy de 75.9%, con un buen f1-score para ambas clases. Es especialmente útil si se quiere minimizar los falsos negativos (fumadores que no son detectados).

# Conclusión

Dado que el objetivo principal es maximizar la capacidad de predecir correctamente a los fumadores, incluso a costa de obtener algunos falsos positivos,   
el modelo más adecuado es el Random Forest (Complejidad reducida).  
  
Este modelo alcanza el mejor recall para la clase de fumadores tanto en entrenamiento (0.94) como en test (0.91), lo que significa que detecta a la mayor cantidad posible de fumadores.   
Aunque puede clasificar por error a algunos no fumadores como fumadores, esto es aceptable si se prioriza la sensibilidad del modelo.