**Resumen de comparación de los modelos**

Se entrenaron y evaluaron cinco variantes de modelos de clasificación: **Random Forest Básico**, **Random Forest con hiperparámetros**, **Random Forest con ajuste de umbral**, **XGBoost estándar** y **XGBoost con el parámetro scale\_pos\_weight ajustado** para mejorar el rendimiento frente al desbalance de clases.

**1. Random Forest (Básico)**

Este modelo muestra una fuerte señal de sobreajuste (overfiting).

* En entrenamiento alcanzó un accuracy perfecto de 1.00, clasificando correctamente todos los casos.
* En test, el accuracy cayó a 79%, con una precisión de 0.72 y un recall de 0.71 para la clase 1 (fumadores).
* Esto indica que aprendió demasiado bien los datos originales, pero no logra generalizar adecuadamente a nuevos casos.

**2. Random Forest (con hiperparámetros)**

Este modelo fue ajustado para reducir la complejidad y mejorar la capacidad de generalización, pero aún así parece tener algo de overfiting.

* En entrenamiento obtuvo un accuracy del 76%, con un recall muy alto (0.94) para fumadores, pero una precisión baja (0.61).
* En test, el accuracy bajó a 73%, con recall de 0.91 y precisión de 0.59 para la clase 1.
* Es eficaz para identificar fumadores, pero comete muchos falsos positivos, clasificando a no fumadores como si lo fueran.

**3. Random Forest (con ajuste de umbral)**

Esta versión modifica el umbral de clasificación para favorecer la detección de la clase minoritaria (fumadores).

* En entrenamiento y test obtuvo un accuracy de 71% y 70% respectivamente.
* Destaca por un recall muy alto (0.96 y 0.95), pero con una precisión baja de 0.56 para la clase 1.
* Es un modelo útil cuando se busca detectar la mayoría de los fumadores, aunque se generen muchos falsos positivos.

**4. XGBoost (modelo estándar)**

Este modelo obtuvo los mejores resultados generales.

* En entrenamiento alcanzó un accuracy del 86%, con buen balance entre precisión (0.80) y recall (0.83) para la clase 1.
* En test, logró un accuracy del 77%, con precisión de 0.68 y recall de 0.71 para fumadores.
* Se destaca por su estabilidad y capacidad de generalización, con un desempeño sólido tanto en datos conocidos como en datos nuevos.

**5. XGBoost (scale\_pos\_weight + hiperparámetros)**

Este modelo fue ajustado para priorizar la detección de la clase positiva en contextos desbalanceados.

* En entrenamiento alcanzó un accuracy del 78%, con recall de 0.90 y precisión de 0.64 para fumadores.
* En test obtuvo un accuracy del 75%, con recall de 0.87 y precisión de 0.61 para la clase 1.
* Es ideal cuando el objetivo es maximizar la detección de fumadores, incluso si eso implica sacrificar algo de precisión.

**Conclusión**

Luego de evaluar el rendimiento de todos los modelos aunque el Random Forest con hiperparametros dio buenos resultados para detectar fumadores, parece estar con un overfiting pequeño, por lo tanto, en su lugar, opté por el modelo XGBoost con scale\_pos\_weight + hiperparámetros. Esta elección la base en su alta capacidad para detectar fumadores, manteniendo una precisión aceptable.

En particular, este modelo alcanzó un recall del 87% en test, lo es una opción bastante confiable cuando el objetivo es no dejar pasar ningún posible caso positivo, aún a costa de generar algunas alertas falsas. Tiene un desempeño consistente entre entrenamiento y test y no se ve overfiting.

Como en un contexto real el modelo podría aplicarse para detectar condiciones de salud, donde es preferible prevenir antes que omitir un caso crítico, esta estrategia prioriza la detección temprana, incluso si eso implica reducir levemente la precisión.