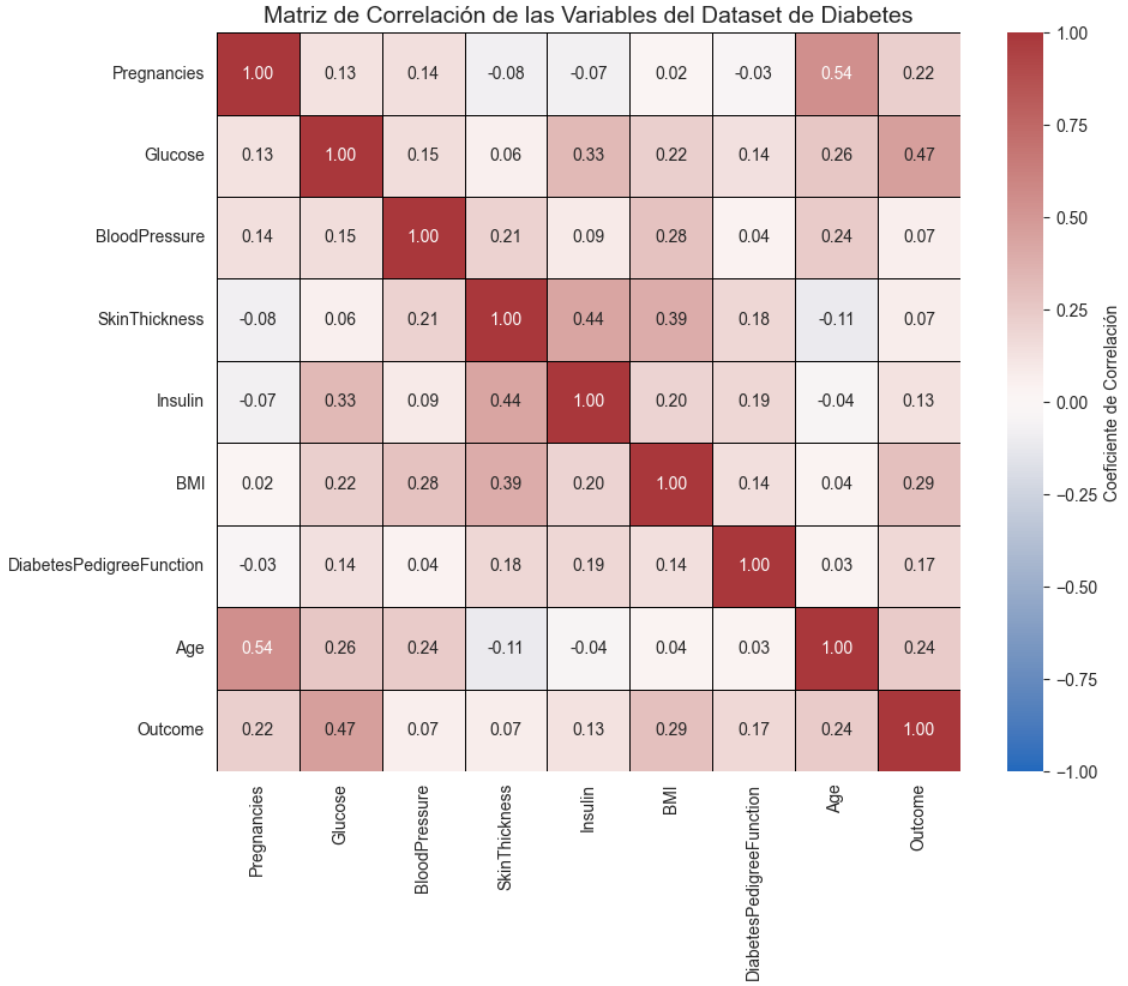
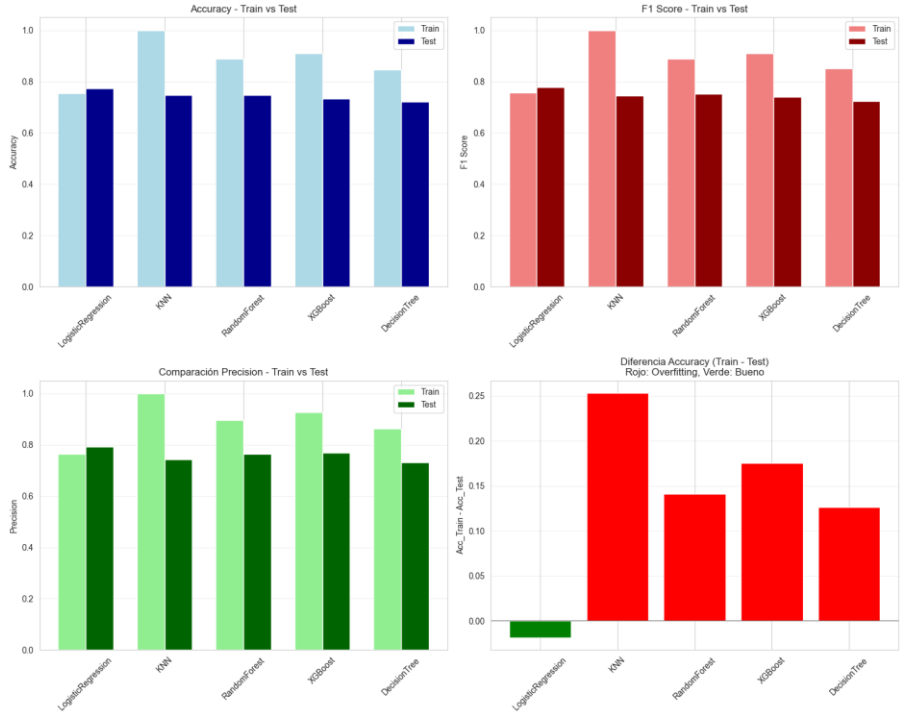
Práctica 2.

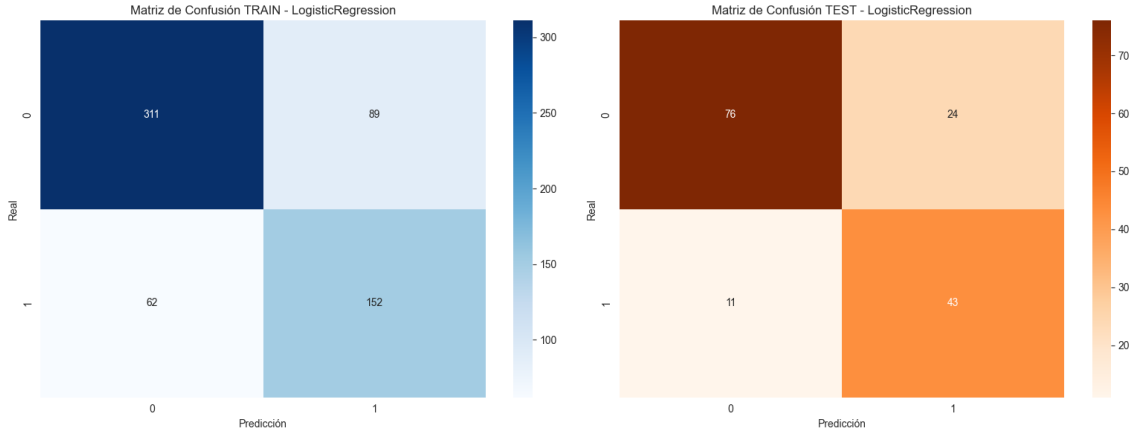
Machine Learning aplicado a dataset sobre Diabetes.



Javier Conejero Rodríguez

Resultados obtenidos.





Para abordar la clasificación de pacientes diabéticos sin alterar la distribución original de los datos (sin resampling), se optó por una estrategia de penalización de costes (Cost-Sensitive Learning).

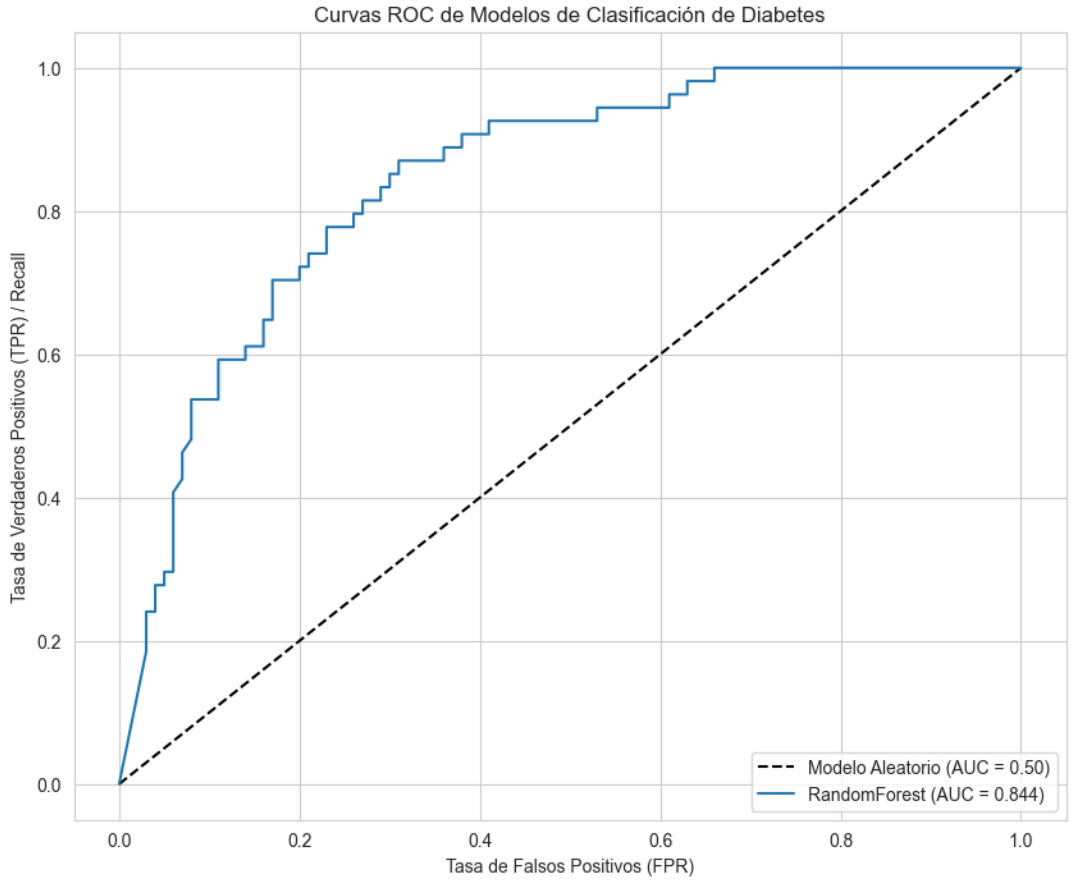
- **Gestión del Desbalance**: Se utilizaron los hiperparámetros de ponderación (class_weight='balanced' o scale_pos_weight) dentro de los algoritmos para otorgar mayor importancia al error en la clase minoritaria (Diabetes).
- **Validación**: Se aplicó Stratified K-Fold (k=5) para garantizar que cada pliegue de validación mantuviera la proporción real de clases, asegurando métricas fiables.

La evaluación de cinco algoritmos distintos revela dos comportamientos opuestos en cuanto a la capacidad de generalización:

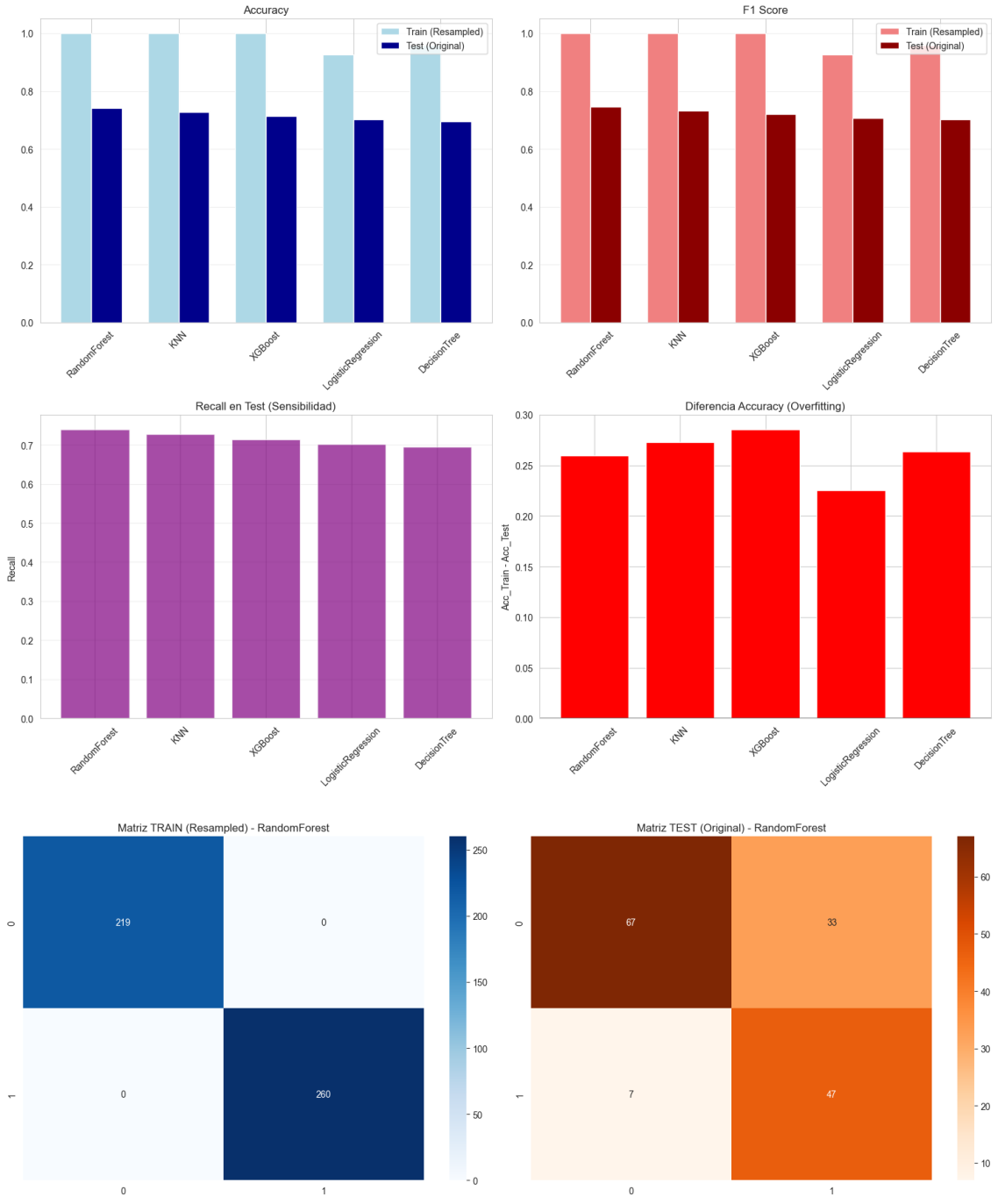
1. **Modelos Basados en Instancias y Árboles (KNN, Random Forest, XGBoost):**
    - Mostraron un sobreajuste (overfitting) severo. Como se observa en el gráfico de *"Diferencia Accuracy"*, modelos como KNN memorizaron el set de entrenamiento (Accuracy ~1.0) pero su rendimiento cayó drásticamente en el set de prueba (>20% de caída).
    - Esto indica que, al intentar ajustarse a la clase minoritaria sin datos sintéticos de apoyo, estos modelos capturaron el ruido estadístico en lugar del patrón real.
2. **Modelos Lineales (Regresión Logística):**
    - Fue el modelo más robusto y estable. Es el único que presenta una diferencia negativa/neutra entre Train y Test (barra verde en el gráfico de diferencias), lo que significa que generaliza perfectamente ante datos nuevos.
    - Alcanzó el mejor equilibrio en F1-Score en Test (~0.77), superando a los modelos más complejos.

Basándonos en su estabilidad, se selecciona la Regresión Logística como el mejor clasificador. Sus métricas detalladas en el conjunto de Test son:

- **Sensibilidad (Recall) - 80%:** Detectó correctamente a 43 de los 54 pacientes diabéticos. Gracias al parámetro class_weight, el modelo priorizó no dejar escapar a los enfermos.
- **Precisión - 64%:** Generó 24 falsos positivos. Esto es un comportamiento esperado en modelos lineales con clases solapadas: para asegurar una alta detección, el modelo debe mover su umbral de decisión hacia una zona "segura", clasificando a algunos sanos dudosos como enfermos.


Curvas ROC de Modelos de Clasificación de Diabetes
Tasa de Verdaderos Positivos (TPR) / Recall
Tasa de Falsos Positivos (FPR)
0.0
0.2
0.4
0.6
0.8
1.0
Modelo Aleatorio (AUC = 0.50)
RandomForest (AUC = 0.844)

Aplicando Resampling.



La aplicación de la técnica híbrida de remuestreo SMOTE-ENN (Synthetic Minority Over-sampling Technique + Edited Nearest Neighbours) ha transformado radicalmente el comportamiento del clasificador Random Forest, priorizando la detección de la clase minoritaria a costa de la precisión global.

A continuación se detallan las métricas obtenidas en el conjunto de Test (datos originales):

- **Sensibilidad (Recall) - Clase 1 (Diabetes): 0.87**

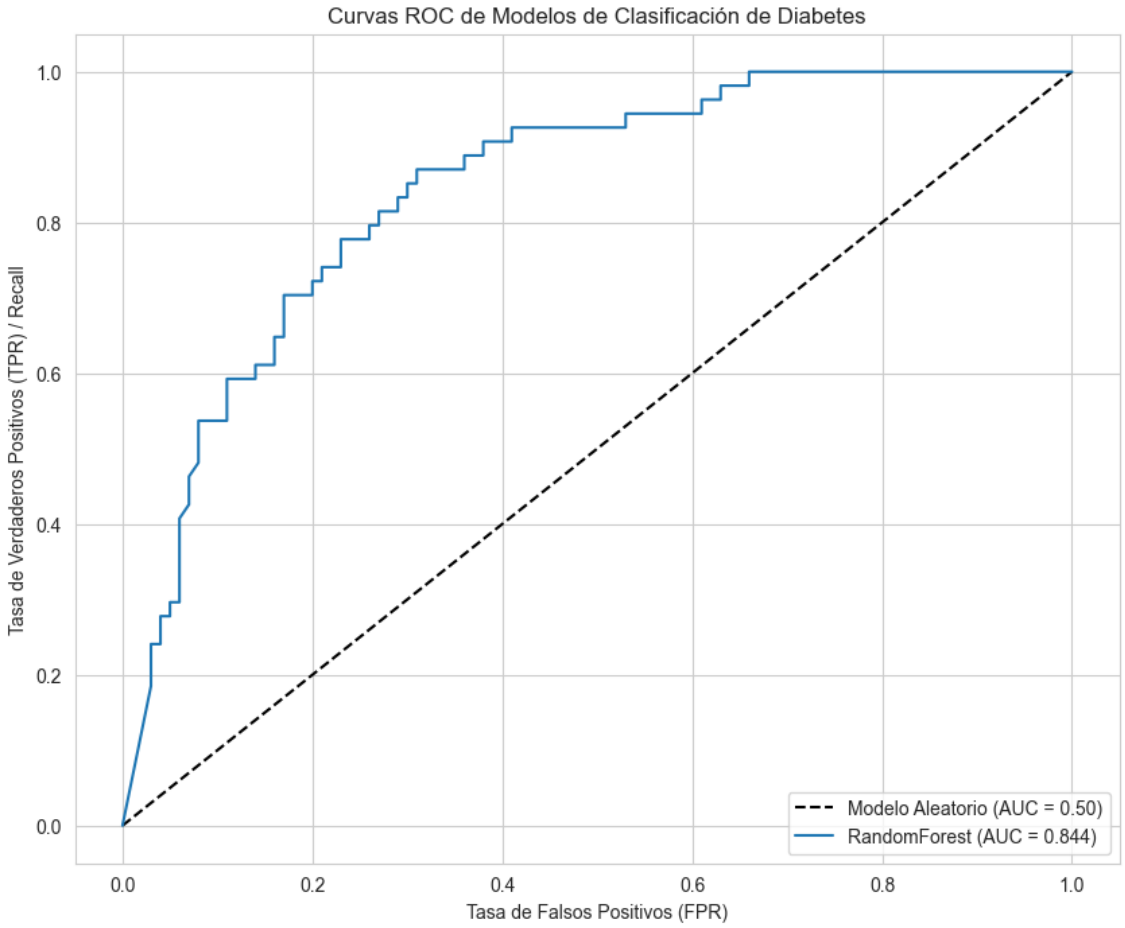
    - Resultado: El modelo identifica correctamente al 87% de los pacientes diabéticos.
    - Causa Técnica: La fase SMOTE generó gran cantidad de ejemplos sintéticos de diabetes, obligando al algoritmo a aprender patrones detallados de esta clase. Simultáneamente, la fase ENN eliminó muestras de la clase mayoritaria (sanos) que invadían la zona de la clase minoritaria, "limpiando" la frontera de decisión para favorecer a la clase positiva.

- **Precisión - Clase 1 (Diabetes): 0.59**
    - Resultado: Solo el 59% de las predicciones positivas fueron correctas. Esto implica una tasa elevada de Falsos Positivos.
    - Causa Técnica: Dado el alto solapamiento visualizado en el PCA, la generación de puntos sintéticos (SMOTE) en zonas densas probablemente invadió el espacio real de los pacientes sanos. Al testear con datos reales, el modelo clasifica como "diabéticos" a muchos pacientes sanos que caen en esas zonas que el modelo aprendió como "territorio diabético" artificialmente.
- **Specificity (Recall de la Clase 0): 0.67**
    - Resultado: El modelo falla al reconocer a los sanos en un 33% de los casos.
    - Interpretación: El modelo se ha vuelto "agresivo" o "paranoico". Ante la duda, tiende a clasificar como enfermo, sacrificando la especificidad.

La implementación de Random Forest combinada con SMOTE-ENN ha generado un modelo de alto cribado (High Sensitivity Screening).

Al modificar la distribución del entrenamiento, se ha logrado romper la barrera de detección habitual, alcanzando un Recall del 87% en la clase Diabetes, superior al de los modelos sin resampling. Sin embargo, este aumento en la sensibilidad conlleva un coste operativo significativo: la Precisión cae al 59%, generando un número elevado de falsos positivos.

Este modelo es la configuración ideal si la prioridad médica es minimizar los falsos negativos (evitar que enfermos se vayan sin diagnóstico), asumiendo que será necesaria una segunda prueba confirmatoria para descartar a los sanos incorrectamente clasificados."


Curvas ROC de Modelos de Clasificación de Diabetes
Tasa de Verdaderos Positivos (TPR) / Recall
Tasa de Falsos Positivos (FPR)
0.0
0.2
0.4
0.6
0.8
1.0
Modelo Aleatorio (AUC = 0.50)
RandomForest (AUC = 0.844)