Proyecto 2

Simulador de Diagnóstico Médico

Rafa, Marcos de Castro Muñoz, Pedro

RESUMEN

01.- Descripción.

Este proyecto tiene como objetivo desarrollar un simulador capaz de predecir enfermedades cardíacas a partir de datos clínicos estructurados.

Entre los datos también contamos con imágenes médicas para predecir posibles enfermedades a través de imágenes.

Además, está integrado con Google Cloud para escalar y desplegar los modelos.

02.- Objetivos.

Diagnóstico temprano de enfermedades cardíacas mediante el estudio y análisis de datos de pacientes reales o similares, haciendo uso de la implementación de modelos de machine learning para ayudarnos en la detección del mismo.

ESTRUCTURA

01.- Exploración de Datos.

- Herramienta usada: ProfileReport (yprofiling). Librería de python que permite generar un html con un estudio en profundidad del contenido del dataset, así como de las relaciones que existen entre columnas.
- Análisis descriptivo de las variables (tipos, distribuciones, correlaciones).
- Observaciones importantes detectadas (valores faltantes, outliers, data-leakage.)

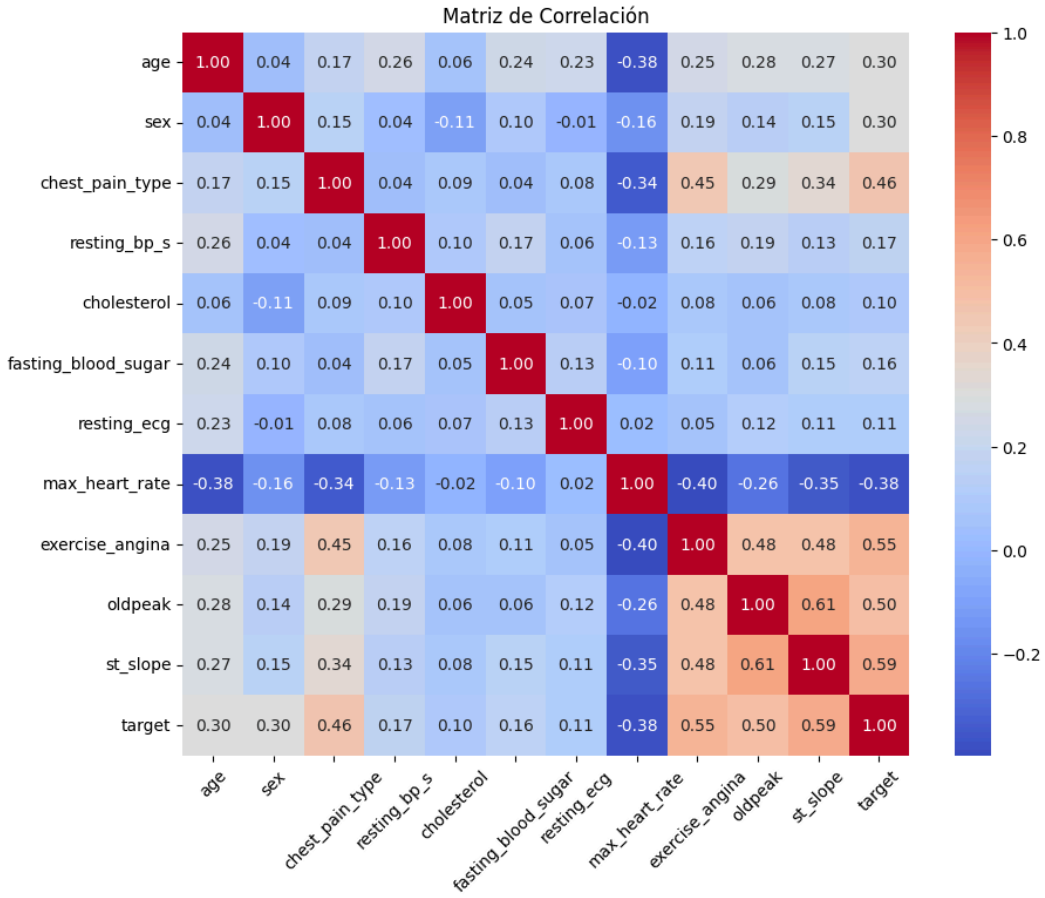
02.- Modelado Predictivo.

2.1. Introducción

Este informe resume el análisis exploratorio, la selección de modelos y el impacto de datos sintéticos en la predicción de enfermedad cardiaca, especialmente considerando el desbalance por sexo.

2.2. Estudio del Dataset Original

A continuación, se muestra la matriz de correlación entre variables:



Se excluyeron las columnas 'exercise_angina', 'oldpeak' y 'st_slope' por posible fuga de información. La siguiente tabla detalla dicha evaluación:

Variable	¿Riesgo de Fuga?	Comentario
`age`, `sex`	❌ No	Datos demográficos seguros.
`chest_pain_type`	⚠️ Bajo	Síntoma subjetivo, pero válido clínicamente como predictor temprano.
`resting_bp_s`, `cholesterol`, `fasting_blood_sugar`, `resting_ecg`	❌ No	Datos de ingreso médico, seguros.
`max_heart_rate`	⚠️ Posible	Obtenido durante una prueba de esfuerzo, pero a veces se considera como parte de una evaluación inicial. Riesgo medio.
`exercise_angina`	✅ Sí	Derivado de una prueba de esfuerzo. Se hace típicamente después de sospechas clínicas.
`oldpeak`	✅ Sí	Resultado de electrocardiograma bajo estrés físico (post-evaluación inicial).
`st_slope`	✅ Sí	Derivado directo de `oldpeak`. Involucra interpretación de prueba de esfuerzo.

Prescindimos de las columnas exercise_angina, oldpeak y st_slope porque:

☐ Riesgo de fuga de información (data leakage)

– Todas ellas describen el comportamiento del paciente durante o tras el ejercicio (angina inducida, depresión del ST), que está muy directamente ligado al diagnóstico de enfermedad cardíaca. Incluirlas habría permitido al modelo «espiar» señales que, en la práctica, solo se conocen después de un test de esfuerzo clínico, inflando artificialmente el rendimiento.

☐ Alta correlación con el target

– En el EDA vimos que estas tres variables presentaban una correlación muy fuerte con la columna target, lo que las hacía redundantes y propensas a causar sobreajuste.

2.3. Modelado Predictivo

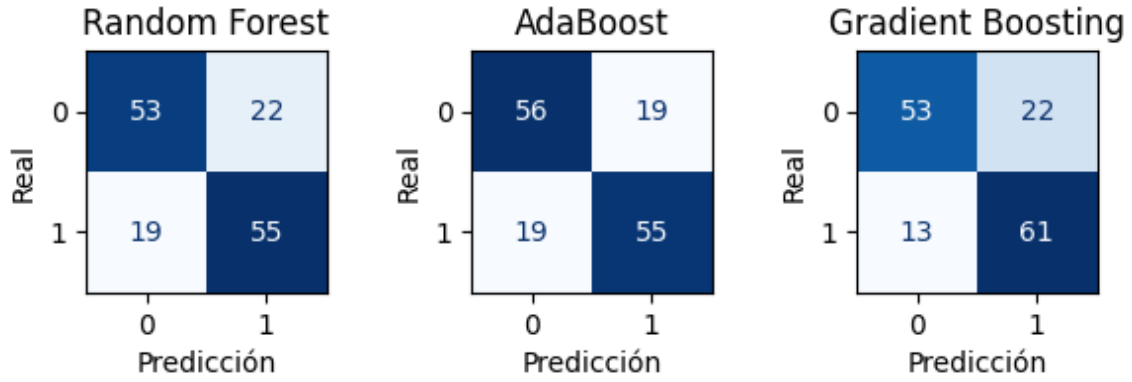
Se probaron varios modelos: Random Forest, AdaBoost y Gradient Boosting. A continuación, se resumen las métricas obtenidas tras optimización con GridSearchCV:

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.725	0.714	0.743	0.728
AdaBoost	0.745	0.743	0.743	0.743
Gradient Boosting	0.765	0.735	0.824	0.777

Modelo	ROC AUC	Accuracy	Precision	Recall	F1-Score
Random Forest	0.8146	0.7248	0.7143	0.7432	0.7285
AdaBoost	0.7835	0.7450	0.7432	0.7432	0.7432
Gradient Boosting	0.8100	0.7651	0.7349	0.8243	0.7771

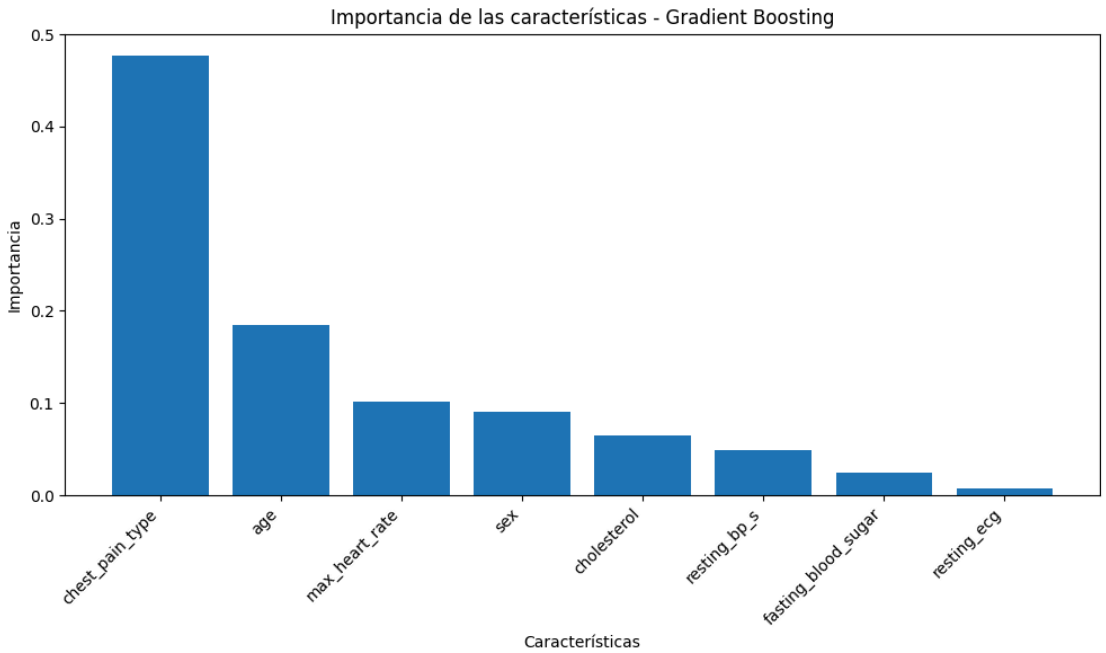
Modelo	Comentario
Gradient Boosting	Mejor F1 y Recall. Modelo más robusto.
AdaBoost	Equilibrado, pero menor recall.
Random Forest	Buen ROC AUC, pero F1 más bajo.

Matrices de confusión de los tres modelos evaluados:



	Predicho: 0 (Sano)	Predicho: 1 (Enfermo)
Real: 0 (Sano)	TP (Verdaderos Negativos)	FP (Falsos Positivos)
Real: 1 (Enfermo)	FN (Falsos Negativos)	TP (Verdaderos Positivos)

Importancia de las variables según el modelo Gradient Boosting:



Evaluar el rendimiento del modelo por sexo

Obserbamos que datos de 'sex' están desbalanceados:

75,6 (1) Hombres

24,4 (0) Mujeres

Evaluar métricas por subgrupo (sex) con Gradient Boosting (mejor modelo)

Resultados por sex del dataset original

```
📋 Resultados para Hombres (1):
Total muestras: 122
Matriz de confusión:
[[36 18]
 [11 57]]
Reporte de clasificación:
              precision    recall  f1-score   support

         0.0       0.77      0.67      0.71        54
         1.0       0.76      0.84      0.80        68

    accuracy                           0.76       122
   macro avg       0.76      0.75      0.76       122
weighted avg       0.76      0.76      0.76       122


📋 Resultados para Mujeres (0):
Total muestras: 27
Matriz de confusión:
[[17  4]
 [ 2  4]]
Reporte de clasificación:
              precision    recall  f1-score   support
...
    accuracy                           0.78        27
   macro avg       0.70      0.74      0.71        27
weighted avg       0.81      0.78      0.79        27
```

Teniendo en cuenta los resultados y los pocos datos de mujeres, procedemos a generar datos sintéticos de mujeres.

Utilizamos MOSTLY AI para generar los datos con el mismo porcentaje de target.

Resultados por sex del dataset ampliado con datos sintéticos (250 mujeres)

```
📋 Resultados para Hombres (1):
Matriz de confusión:
[[34  9]
 [ 8 63]]
Reporte de clasificación:
              precision    recall  f1-score   support

         0.0       0.81      0.79      0.80        43
         1.0       0.88      0.89      0.88        71

    accuracy                           0.85       114
   macro avg       0.84      0.84      0.84       114
weighted avg       0.85      0.85      0.85       114


📋 Resultados para Mujeres (0):
Matriz de confusión:
[[64  1]
 [10 10]]
Reporte de clasificación:
              precision    recall  f1-score   support

         0.0       0.86      0.98      0.92        65
         1.0       0.91      0.50      0.65        20
...
    accuracy                           0.87        85
   macro avg       0.89      0.74      0.78        85
weighted avg       0.88      0.87      0.86        85
```

Los resultados con 250 datos sintéticos de mujeres para balancear el dataframe df_heart mejoran las predicciones tanto de mujeres cómo de hombres.

Comparativas

1. Comparativa por sexo

2. Comparativa global (sin dividir por sexo)

1. Comparativa por sexo

📈 Mejora notable en mujeres, sobre todo en f1-score (+0.08) y accuracy (+0.09).

👨 En hombres también se observa mejora, posiblemente por regularización o mayor equilibrio general.

Hombres (sex = 1)

Métrica	Dataset original (`df_heart`)	Dataset aumentado (`df_augmented`)
Accuracy	0.76	**0.85** ✅
Precision (1)	0.76	**0.88** ✅
Recall (1)	0.84	0.89 ✅
F1-Score (1)	0.80	**0.88** ✅
Total muestras	122	114

Mujeres (sex = 0)

Métrica	Dataset original (`df_heart`)	Dataset aumentado (`df_augmented`)
Accuracy	0.78	**0.87** ✅
Precision (1)	0.50	**0.91** ✅
Recall (1)	0.67	0.50 ❌ (ligera caída)
F1-Score (1)	0.57	**0.65** ✅
Total muestras	27	85

📘 2. Comparativa global (sin dividir por sexo)

🔍 Suben todas las métricas globales, demostrando que:

El dataset extendido mejora la generalización.

Los nuevos datos no introducen ruido ni sesgo.

Dataset	Accuracy	Precision	Recall	F1-Score
Original	0.78	0.72	0.76	0.74
Con 250 sintéticos	0.86	0.87	0.77	0.81

INFRAESTRUCTURA

01.- Google Cloud.

Para el almacenamiento de datos hemos utilizado un bucket de google cloud, que nos permite guardar grandes cantidades de datos facilitando la escalabilidad del proyecto sin incrementar la complejidad del mismo.

02.- Estructura Cloud.

La estructura definida es la siguiente:

Un bucket con acceso mediante la creación de un service account, utilizamos la clave json que habilita este usuario, con rol habilitado de administración de almacenamiento para poder acceder a los datos subidos, subir nuevos datos y descargarlos, o eliminarlos en caso de que fuese necesario.

Conclusión y Próximos Pasos

01.- Conclusión.

Hemos tratado los datos, comprobar que los datos estuvieran equilibrados y en caso de no estarlos, añadir para balancear. Hemos probado diferentes modelos de Machine Learning y obtenido sus métricas para poder elegir que tan bien resuelven nuestro problema, lo que como hemos explicado antes es el GradientBoosting.

Gradient Boosting es el modelo elegido porque, tras la optimización de hiperparámetros, obtuvo el mejor equilibrio entre precisión y cobertura de casos positivos:

F1-Score más alto (0.777), superando a Random Forest (0.728) y AdaBoost (0.743).

Recall líder (0.824), esencial para minimizar los falsos negativos en un contexto médico, por delante de Random Forest (0.743) y AdaBoost (0.743).

ROC AUC competitivo (0.810), casi igual al de Random Forest (0.815) y muy superior a AdaBoost (0.784).

Este perfil—alta sensibilidad sin sacrificar precisión—garantiza un modelo fiable para la detección temprana de enfermedad cardíaca.

Añadir 250 registros sintéticos de mujeres (con la misma proporción 75%/25% de target) mejora significativamente el rendimiento del modelo en mujeres, donde antes mostraba peores métricas, sin perjudicar —e incluso mejorando— los resultados en hombres y a nivel general.

02.- Próximos Pasos.

Debido a los pocos datos (imágenes) encontradas clasificadas, no hemos podido añadir la posibilidad de detección de enfermedades a través de imágenes pero lo añado aquí ya que está en los planes del proyecto.

Anexos

Data:

ProyectoReconocimientoEnfermedad/data/raw/heart-disease-dataset.csv at main · GrupoProyecto2IABD/ProyectoReconocimientoEnfermedad

Repo:
GrupoProyecto2IABD/ProyectoReconocimientoEnfermedad