DETECCIÓN DE ENFERMEDADES CARDIOVASCULARES

1. Objetivo del proyecto

Las enfermedades del corazón son una de las principales causas de muerte y problemas de salud en el mundo, lo que genera una gran presión en hospitales como La Fe. El mayor desafío es detectar a tiempo a los pacientes en riesgo. Si un paciente con ECV no es identificado, las consecuencias pueden ser muy graves y costosas. Por eso, este proyecto se enfoca en crear una herramienta que minimice la posibilidad de 'falsos negativos' (no detectar a alguien que sí está enfermo), priorizando la detección exhaustiva para la seguridad del paciente y la eficiencia de La Fe.

2. Dataset y Fuentes de Datos

Para entrenar nuestro sistema, usamos un conjunto de datos público de Kaggle llamado "Cardiovascular Disease Dataset", con 70.000 registros de pacientes. Este dataset contiene información básica de salud y hábitos de vida, como edad, peso, tensión arterial, colesterol, y si fuman o hacen ejercicio. La información es similar a la que se obtiene en una consulta médica normal, lo que lo hace muy práctico para La Fe.

Es importante destacar que el dataset está bastante equilibrado (aproximadamente la mitad de los pacientes tienen ECV y la otra mitad no). Esto es bueno porque nos ayuda a construir un modelo imparcial y a obtener resultados de evaluación fiables.

3. Análisis de Datos Clave

Analizamos a fondo los datos para entender qué factores son más importantes para predecir las ECV:

- Factores de Riesgo Clave: Identificamos que la edad, la presión arterial, el colesterol y el índice de masa corporal (IMC) son los factores que más influyen en la predicción de ECV. Esto permite a La Fe concentrar sus esfuerzos de prevención en estos aspectos.

4. Preparación de Datos y Construcción del Modelo

Antes de entrenar el sistema, preparamos los datos cuidadosamente:

- Dividimos los datos: Separamos el 80% para entrenar el modelo y el 20% para probarlo con información que no había visto antes.
- Normalizamos los datos: Ajustamos los valores numéricos para que el modelo aprendiera de forma más eficiente.
- Optimización: Usamos técnicas avanzadas (GridSearchCV y RandomizedSearchCV con validación cruzada) para encontrar la mejor configuración para nuestro modelo, asegurando que fuera robusto y preciso.

5. Evaluación del Rendimiento del Modelo

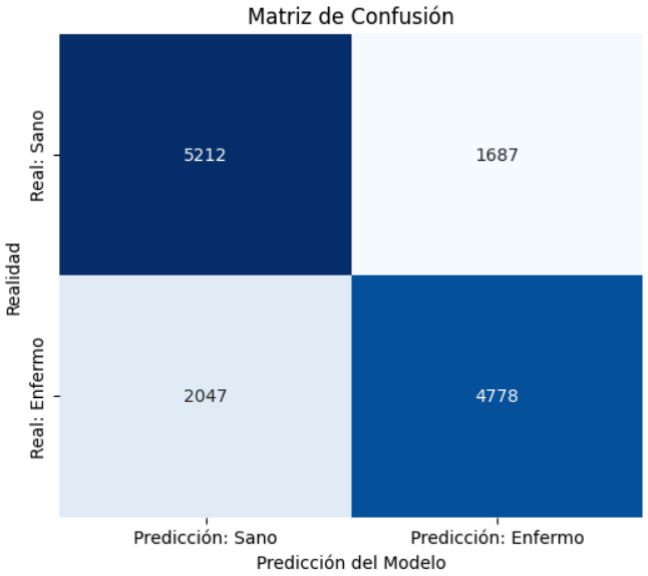
Para medir qué tan bien funciona el modelo, usamos varias métricas, siendo la más importante la Sensibilidad (o Recall):

- **Precisión (Accuracy):** El modelo acierta el 72.79% de las veces en general.
- **Sensibilidad (Recall):** Detecta correctamente al 70.01% de los pacientes que ***realmente* tienen ECV.** Esta es nuestra prioridad para La Fe, para no dejar pasar casos importantes.
- Otras métricas como la Precisión, F1 Score y ROC AUC también fueron consideradas para tener una visión completa del rendimiento.

6. Resultados Clave: El Modelo Elegido y Su Impacto

Analizamos varios modelos (Regresión Logística, Random Forest, AdaBoost, XGBoost), y el que mejor se adaptó a nuestro objetivo de alta detección fue Random Forest. Fue optimizado para asegurar esa alta capacidad de identificar a los enfermos.

6.1. Impacto de la Detección: La Matriz de Confusión



La matriz de confusión nos muestra el rendimiento del modelo en términos de pacientes reales:

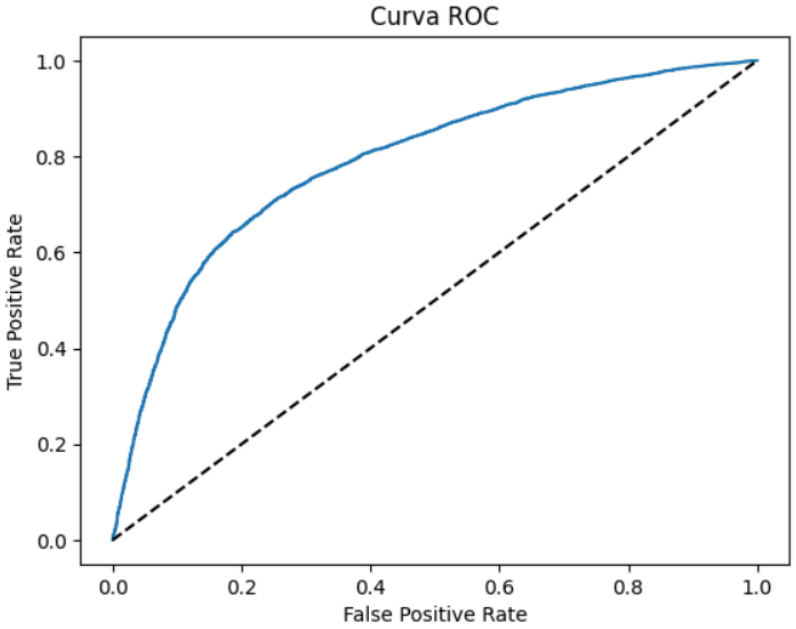
- **Casos Detectados Correctamente (Verdaderos Positivos):** Nuestro modelo identifica a 4778 pacientes que *realmente tienen* ECV. Para La Fe, esto significa oportunidades directas de intervenir a tiempo, lo que puede salvar vidas y evitar tratamientos más caros y complejos en el futuro.
- **Casos No Detectados (Falsos Negativos):** Solo 2047 pacientes que *sí tenían* ECV no fueron detectados. Hemos trabajado mucho para que este número sea lo más

bajo posible, ya que para La Fe, no detectar a un paciente enfermo es el riesgo más alto.

- **Falsos Positivos:** El modelo predice que 4863 pacientes sanos tienen ECV. Esto podría implicar algunas pruebas adicionales, pero es un coste mucho menor comparado con no detectar un caso real de ECV.

Esta matriz demuestra que el modelo es muy adecuado para identificar pacientes en fases iniciales o de cribado, donde detectar la mayoría de los casos es fundamental.

7.2. Fiabilidad General del Modelo: La Curva ROC y AUC



La Curva ROC y su valor AUC (0.7895) nos indican la fiabilidad general de nuestro modelo. Un valor AUC cercano a 1.0 significa que el modelo es muy bueno diferenciando entre pacientes con y sin ECV.

Aunque otros modelos pudieron tener una AUC ligeramente superior, nuestra decisión final se basó en la prioridad de maximizar la Sensibilidad (Recall). Esto asegura que el modelo esté ajustado para minimizar los 'falsos negativos' más críticos, que son los que tienen mayor impacto en la salud del paciente y en los recursos de La Fe. En resumen, el modelo es una herramienta robusta y confiable para ayudar en la toma de decisiones clínicas.

8. Conclusiones y Valor Estratégico para La Fe

El modelo Random Forest desarrollado es una herramienta segura, robusta y práctica para la detección temprana de ECV en el Hospital La Fe.

- Su alta sensibilidad garantiza que se detecten la mayoría de los casos de ECV, lo que reduce el riesgo para los pacientes y los costes hospitalarios futuros.
- El modelo es fácil de entender en sus decisiones. Al basarse en "árboles de decisión", sus resultados pueden ser explicados a médicos y personal no técnico, lo que facilita su confianza y adopción en La Fe.
- Este sistema permitirá a La Fe optimizar sus recursos, mejorar los resultados de salud y reforzar su posición como un hospital líder en innovación médica.