# Predicción de Enfermedades Cardíacas mediante Aprendizaje Automático

Federico Nicolas Trujillo

***Abstract*—Resumen— Se presenta *CardioPredict*, un proyecto de aprendizaje automático enfocado en la predicción de enfermedades cardíacas. Se entrenaron y compararon modelos de clasificación sobre tres conjuntos de datos abiertos (UCI Heart Disease, Heart Failure Prediction de Kaggle y Framingham Heart Study). La metodología incluyó limpieza de datos, codificación de variables categóricas y entrenamiento de algoritmos de regresión logística, bosque aleatorio y XGBoost. Los resultados muestran que el modelo de bosque aleatorio alcanzó la mayor exactitud (hasta un 90% en el dataset de Kaggle), mientras que la regresión logística destacó en escenarios desequilibrados al detectar más casos positivos. Esto demuestra el potencial de las técnicas de aprendizaje automático para apoyar la detección temprana de riesgos cardíacos, aunque se requiere considerar el balance de clases e interpretabilidad para su aplicación clínica.**

***Index Terms*—Aprendizaje automático; Enfermedad cardíaca; Predicción; Regresión logística; Random Forest; XGBoost**

## I. Introducción

Las enfermedades cardiovasculares (ECV) son la principal causa de muerte a nivel mundial, cobrando aproximadamente 17.9 millones de vidas cada año según la Organización Mundial de la Salud (OMS). La detección temprana de enfermedades cardíacas es fundamental para implementar intervenciones preventivas y reducir la mortalidad. En la última década, se han explorado modelos de *Machine Learning* para predecir riesgos cardíacos con resultados prometedores. Estos modelos pueden analizar grandes volúmenes de datos médicos y encontrar patrones complejos que escapan a los métodos tradicionales, mejorando la precisión diagnóstica.

El presente trabajo introduce *CardioPredict*, un sistema que aplica técnicas de aprendizaje automático para la predicción de enfermedades cardíacas utilizando datos de múltiples fuentes públicas. A diferencia de estudios previos que se enfocan en un solo conjunto de datos o algoritmo, *CardioPredict* compara el desempeño de varios modelos de clasificación (regresión logística, *Random Forest* y XGBoost) en tres *datasets* distintos. El objetivo es identificar qué modelo ofrece mejor rendimiento en cada contexto y analizar las razones, aportando una perspectiva sobre cómo las características del conjunto de datos (tamaño, distribución de clases) influyen en la eficacia de cada algoritmo. Los resultados obtenidos muestran altos niveles de exactitud y área bajo la curva ROC (AUC) en la predicción de riesgo cardíaco, destacando la viabilidad de estas herramientas para apoyar la toma de decisiones médicas.

## II. Conjuntos de Datos

Se emplearon tres conjuntos de datos públicos en los experimentos, con características y objetivos resumidos a continuación:

**UCI Heart Disease:** Proviene del repositorio de aprendizaje automático de la UCI e incluye 303 muestras con 14 atributos clínicos (edad, sexo, presión arterial en reposo, colesterol, etc.). El objetivo es diagnosticar la presencia de enfermedad cardíaca (angina) en pacientes, codificada como variable binaria (1 = presencia de enfermedad, 0 = ausencia). Es un *dataset* ampliamente utilizado en la literatura para evaluar algoritmos de clasificación en medicina.

**Kaggle Heart Failure:** Conjunto de datos compilado a partir de cinco fuentes distintas (incluyendo Cleveland, Statlog y otras), unificadas en 11 características comunes, lo que lo convierte en uno de los mayores *datasets* disponibles sobre enfermedad cardíaca con 918 registros. Incluye variables categóricas como tipo de dolor de pecho, resultados de ECG, angina inducida por ejercicio, etc., y variables numéricas como frecuencia cardíaca máxima y nivel de colesterol. La variable objetivo es *HeartDisease* (1 = presencia de enfermedad cardiovascular, 0 = normal). No presenta valores perdidos en sus campos, facilitando su uso directo.

**Framingham Heart Study:** Conjunto de datos derivado del estudio longitudinal de Framingham, disponible en Kaggle, que contiene datos de 4,238 pacientes con 15 atributos de factores de riesgo (edad, sexo, índice de masa corporal, presión arterial, hábitos de tabaco, diabetes, entre otros) El objetivo es predecir el riesgo de desarrollar enfermedad coronaria (CHD) a 10 años, indicado por una etiqueta binaria (*TenYearCHD*). Este *dataset* presenta un fuerte desequilibrio de clases: aproximadamente solo un 15% de los registros están etiquetados como positivos (evento de CHD en 10 años), lo cual representa un desafío para los modelos de clasificación en términos de sensibilidad.

## III. Metodología

### A. Preparación de los Datos

Cada conjunto de datos se dividió en subconjuntos de entrenamiento (70%) y prueba (30%), conservando la proporción de clases para asegurar una evaluación justa. Previamente, se realizó limpieza básica: en el caso de Framingham, se eliminaron o imputaron registros con valores faltantes en variables como colesterol o presión arterial (si los había), y en los demás conjuntos no fue necesario imputar datos debido a la ausencia de nulos significativos. Las variables categóricas (p. ej., tipo de dolor de pecho, resultados de ECG, tipo de angina) fueron codificadas mediante *one-hot encoding* o etiquetas numéricas según conveniencia, de modo que los modelos pudieran procesarlas. Adicionalmente, las

características numéricas fueron normalizadas (escalado min-max) antes de entrenar la regresión logística, con el fin de evitar que diferencias de escala afecten su convergencia.

### B. Modelos de Clasificación

Se evaluaron tres algoritmos de *machine learning* supervisado para la clasificación binaria:

- **Regresión Logística (LR):** Modelo lineal generalizado que estima la probabilidad de la clase positiva mediante una función sigmoide. Se utilizó con regularización $L_2$ por defecto para evitar sobreajuste. La regresión logística proporciona una base de comparación y es interpretable via coeficientes asociados a cada factor de riesgo.

- **Bosque Aleatorio (RF):** Ensamble de múltiples árboles de decisión entrenados con *bagging*. Se entrenó un *Random Forest* con 100 árboles y profundidad máxima libre (determinada por los datos) usando el criterio de Gini para medidas de impureza. Este método captura relaciones no lineales e interacciones entre variables de forma efectiva, a costa de menor interpretabilidad individual.

- **XGBoost (XGB):** Implementación optimizada de *boosting* de árboles (árboles de decisión agregados secuencialmente) que minimiza un objetivo de pérdida regularizado. Se configuró con un número de estimadores inicial (100 árboles) y una tasa de aprendizaje estándar (0.1). XGBoost tiende a lograr alto rendimiento en muchos problemas de clasificación gracias a su reducción de sesgo en el entrenamiento, aunque requiere ajuste cuidadoso de hiperparámetros para evitar sobreajuste.

Todos los modelos se entrenaron utilizando Python (bibliotecas scikit-learn y XGBoost) con los hiperparámetros mencionados o por defecto en caso no especificado. No se aplicó ajuste extensivo de hiperparámetros dado el alcance del proyecto; en su lugar, se utilizó una configuración razonable estándar y se evaluó el desempeño en el conjunto de prueba para comparar de forma consistente las capacidades predictivas de cada algoritmo en cada *dataset*.

### C. Evaluación e Interpretación

Para medir el rendimiento de los modelos se emplearon varias métricas de clasificación: *accuracy* o exactitud (proporción de aciertos sobre el total), precisión positiva (valor predictivo positivo), sensibilidad (*recall* o tasa de verdaderos positivos), puntaje $F_1$ (media armónica de precisión y sensibilidad) y el área bajo la curva ROC (AUC). Estas métricas permiten un análisis integral, especialmente en escenarios con clases desequilibradas donde la exactitud por sí sola puede ser engañosa. Por ejemplo, la sensibilidad es crucial en la predicción de enfermedades (minimizar falsos negativos), mientras que la precisión refleja la confiabilidad de las detecciones positivas.

Se construyeron matrices de confusión para cada modelo con el fin de examinar el detalle de aciertos y errores por clase. Adicionalmente, con el objetivo de interpretar las predicciones de los modelos de tipo *caja negra* (en particular, los basados en árboles), se empleó SHAP (*SHapley Additive exPlanations*). Esta técnica atribuye a cada característica un valor que representa su contribución a la predicción, proporcionando una explicación consistente con la teoría de valores de Shapley. Se generaron gráficos de resumen con SHAP para el mejor modelo, identificando qué variables tuvieron mayor influencia en la decisión final.

## IV. Resultados

En la Tabla I se resumen los resultados de las métricas para cada modelo en los tres conjuntos de datos. En el *dataset* UCI, la regresión logística obtuvo la mejor exactitud (85.7%) y AUC (0.938), superando ligeramente al bosque aleatorio (84.6% de exactitud, AUC 0.931). XGBoost alcanzó un rendimiento algo menor en UCI (80.2% exactitud). En el conjunto de Kaggle, el modelo de bosque aleatorio destacó con un 90.2% de exactitud y el valor $F_1$ más alto (0.913), seguido de cerca por la regresión logística (88.4% exactitud) y XGBoost (87.7%). En este caso, Random Forest mostró una mayor capacidad predictiva, reflejada también en la curva ROC (AUC = 0.946 frente a 0.932 de la regresión logística) gracias a su habilidad para capturar relaciones no lineales en un conjunto de datos relativamente grande.

En contraste, para el *dataset* de Framingham, la métrica de exactitud fue más alta con Random Forest (84.7%), pero a costa de una sensibilidad extremadamente baja (solo 2%), lo que indica que prácticamente no detectó los casos positivos (predijo casi todo como "no enfermo" debido al desequilibrio de clases). De hecho, XGBoost presentó un comportamiento similar (82.5% exactitud, 9.3% de sensibilidad). La regresión logística, aunque tuvo una exactitud más modesta (67.2%), logró identificar alrededor del 61% de los pacientes que desarrollaron enfermedad (sensibilidad mucho mayor) y obtuvo el $F_1$ más alto (0.36) en Framingham. Esto sugiere que el modelo lineal fue más efectivo en manejar el desequilibrio al no sesgarse completamente hacia la clase mayoritaria, a diferencia de los modelos de árboles que optimizaron la exactitud global a expensas de no captar los positivos. En aplicaciones médicas, este resultado es relevante puesto que un clasificador con alta sensibilidad puede ser preferible para no pasar por alto pacientes en riesgo, incluso si implica cierta reducción de la precisión o mayor falsa alarma.

TABLE I
Desempeño de los Modelos en Cada Conjunto de Datos

| Dataset | Modelo | Acc | Prec | Rec | F1 | AUC |
|---|---|---|---|---|---|---|
| UCI | Logist. | 0.857 | 0.837 | 0.857 | 0.847 | 0.938 |
| UCI | RandomF | 0.846 | 0.804 | 0.881 | 0.841 | 0.931 |
| UCI | XGBoost | 0.802 | 0.773 | 0.810 | 0.791 | 0.894 |
| Kaggle | Logist. | 0.884 | 0.906 | 0.882 | 0.894 | 0.932 |
| Kaggle | RandomF | 0.902 | 0.899 | 0.928 | 0.913 | 0.946 |
| Kaggle | XGBoost | 0.877 | 0.899 | 0.876 | 0.887 | 0.935 |
| Fram. | Logist. | 0.672 | 0.257 | 0.611 | 0.361 | 0.700 |
| Fram. | RandomF | 0.847 | 0.444 | 0.021 | 0.040 | 0.669 |
| Fram. | XGBoost | 0.825 | 0.277 | 0.093 | 0.140 | 0.633 |

En la Figura 1 se muestra la matriz de confusión del modelo Random Forest para el conjunto de datos de Kaggle (Heart Failure). Se observa una alta proporción de verdaderos

positivos y verdaderos negativos, concordante con su elevada exactitud; los pocos errores se concentran en falsos negativos levemente más que falsos positivos (lo cual es favorable en contextos donde es preferible errar por sobre-detección que no detectar). La Figura 2 presenta las curvas ROC de los tres modelos en este mismo conjunto: la curva del bosque aleatorio alcanza la mayor área (más cercana a la esquina superior izquierda), seguida muy de cerca por XGBoost y la regresión logística, lo que indica que todos logran buen poder discriminatorio con diferencias marginales. Finalmente, la Figura 3 muestra un gráfico de valores SHAP para el modelo XGBoost (dataset de Kaggle), evidenciando la importancia de cada característica en la predicción. Entre las variables más influyentes según SHAP se encuentran el tipo de dolor de pecho, la depresión del ST inducida por ejercicio (*Oldpeak*) y la pendiente del segmento ST, las cuales tienen efectos significativos en la probabilidad de enfermedad. Esto coincide con el conocimiento médico, donde anomalías en el electrocardiograma de esfuerzo y ciertos tipos de dolor torácico son indicadores críticos de cardiopatía.

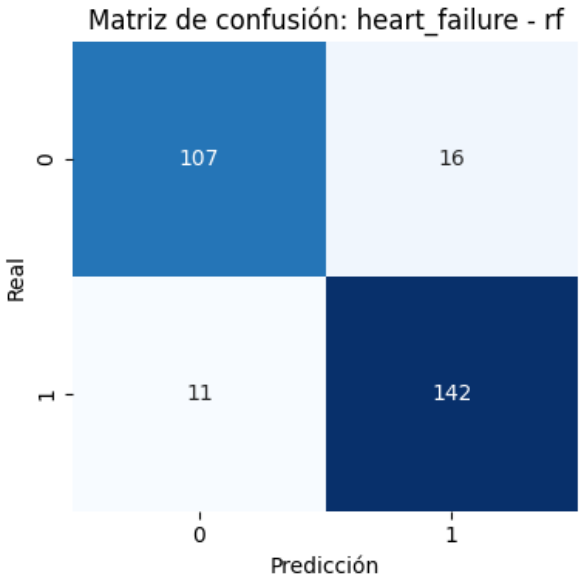


Fig. 1. Matriz de confusión del modelo *Random Forest* en el dataset de Kaggle (Heart Failure). En la diagonal se ubican las predicciones correctas: 142 verdaderos positivos y 107 verdaderos negativos en este caso, con muy pocos falsos positivos (5) y falsos negativos (16).

## V. Discusión

Los experimentos muestran que no existe un modelo único que sea óptimo para todos los conjuntos de datos; el desempeño relativo depende de las características de cada *dataset*. En el caso del conjunto de datos de Kaggle (el mayor y con distribución relativamente balanceada), el bosque aleatorio sobresalió en métricas globales. Su capacidad para aprovechar relaciones no lineales e interacciones entre variables le permitió lograr la mayor exactitud y $F_1$. Además, dado el tamaño de este *dataset*, los modelos complejos como

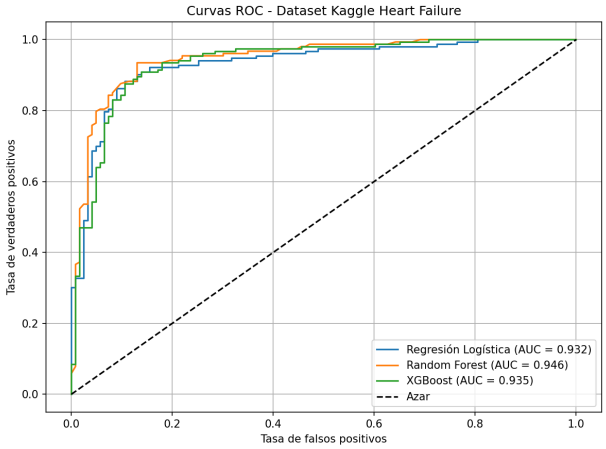


Fig. 2. Curvas ROC de los tres modelos evaluados en el conjunto de datos de Kaggle. La curva del Random Forest (RF) domina ligeramente sobre las de XGBoost (XGB) y Regresión Logística (LR), mostrando una mayor sensibilidad para cualquier nivel de especificidad. El valor de AUC para RF fue 0.946, comparado con 0.935 (XGB) y 0.932 (LR).

RF y XGBoost pudieron entrenar patrones más elaborados sin incurrir en tanto sobreajuste, algo que quizás explica su ventaja ligera sobre la regresión logística.

Para el conjunto UCI (más pequeño, 303 muestras), la diferencia entre modelos fue menor. La regresión logística tuvo un rendimiento ligeramente superior en AUC y estuvo a la par en $F_1$ con Random Forest. Esto sugiere que, con pocos datos, un modelo más sencillo como la regresión lineal generalizada puede generalizar tan bien como modelos no lineales más complejos, evitando sobreajustar ruidos. En UCI, las relaciones pueden ser lo suficientemente lineales o simples para que un modelo lineal capture la señal principal (de hecho, estudios previos con ese *dataset* han encontrado buenos resultados con algoritmos relativamente simples).

El caso de Framingham resalta la importancia de considerar el desequilibrio de clases al seleccionar un modelo. Aquí, los algoritmos basados en árboles optimizaron la exactitud global prediciendo casi siempre la clase negativa (ningún evento de CHD), dado que esa es la mayoría abrumadora. Esto llevó a altas tasas de acierto global pero a una tasa de falsos negativos inaceptablemente alta desde el punto de vista clínico. La regresión logística, en cambio, al modelar probabilidades, permitió un umbral implícito que resultó en más verdaderos positivos. Aunque su exactitud fue menor, captó muchos más casos de enfermedad. Este resultado indica que, sin técnicas adicionales (p.ej., re-muestreo, ajuste de umbral o penalización de clases), los modelos de árboles pueden ser menos adecuados en escenarios desequilibrados si el interés principal es la detección de la clase minoritaria. Un ajuste posible sería incorporar un peso de clase mayor para los positivos durante el entrenamiento de RF y XGBoost, o usar técnicas de sobremuestreo de la clase minoritaria, para mejorar su sensibilidad en futuros trabajos.

En términos de interpretabilidad, la regresión logística ofrece coeficientes directamente interpretables que indican el

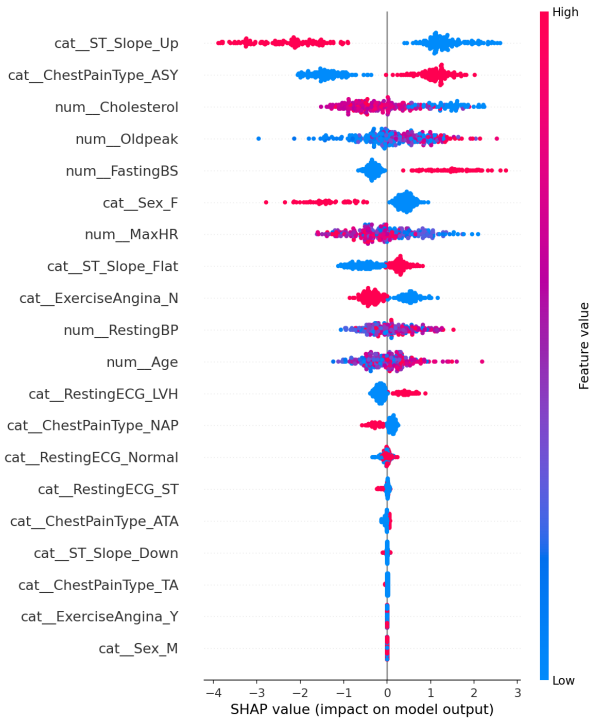


Fig. 3. Resumen de valores SHAP para las características del modelo XGBoost (dataset de Kaggle). Cada punto representa una observación, y el color indica si el valor de la característica es alto (rojo) o bajo (azul). Las variables se ordenan por importancia promedio: aquí *ChestPainType* (tipo de dolor de pecho), *Oldpeak* (depresión ST) y *MaxHR* (frecuencia cardíaca máxima) figuran entre las más determinantes. Se aprecia que, por ejemplo, ciertos tipos de dolor de pecho (valor alto en la categoría angina típica) y mayores valores de *Oldpeak* aumentan significativamente la probabilidad predicha de enfermedad.

efecto de cada variable en la probabilidad de enfermedad. No obstante, los modelos de mejor desempeño fueron los de ensamblaje de árboles, que son esencialmente *cajas negras*. Aquí es donde las herramientas como SHAP resultan valiosas: el análisis SHAP permitió confirmar que las variables identificadas como importantes (dolor de pecho, indicadores del ECG de esfuerzo, frecuencia cardíaca máxima, etc.) concuerdan con el conocimiento médico existente, lo cual brinda confianza en que el modelo está aprendiendo relaciones válidas y no artefactos espúreos. Esta interpretabilidad es crucial para la aceptación de modelos de ML en entornos clínicos, ya que los profesionales de la salud necesitan entender y confiar en las razones detrás de una predicción.

En suma, *CardioPredict* demostró la viabilidad de predecir riesgos de enfermedad cardíaca con alta precisión mediante ML, pero también evidenció que la elección del modelo debe considerar el contexto de los datos. Modelos complejos como Random Forest y XGBoost tienden a rendir mejor con muchos datos y patrones no lineales, mientras que en conjuntos pequeños o altamente desequilibrados, un modelo más simple o estrategias específicas para balance de clases pueden ofrecer ventajas significativas.

## VI. Conclusiones

Este trabajo presentó una comparación de tres técnicas de aprendizaje automático aplicadas a la predicción de enfermedades cardíacas en distintos conjuntos de datos abiertos. Se comprobó que los modelos ensamble de árboles (Random Forest, XGBoost) alcanzan altos niveles de exactitud y AUC en un *dataset* amplio y heterogéneo, superando ligeramente a un modelo lineal, mientras que la regresión logística mostró ser más robusta en escenarios con datos escasos o desbalanceados al mantener una sensibilidad superior. En general, Random Forest fue el modelo con mejor desempeño global en las condiciones evaluadas, logrando hasta 90% de exactitud, seguido de cerca por XGBoost. No obstante, la preferencia por un modelo u otro debe basarse en el criterio clínico: para tamizajes preventivos podría priorizarse aquel que detecte más casos (aunque implique más falsos positivos), mientras que para evitar alarmas innecesarias podría interesar maximizar la precisión.

Los hallazgos resaltan la importancia de abordar el balance de clases al entrenar modelos para predicción médica. Futuros desarrollos de *CardioPredict* incluirán la incorporación de técnicas de re-balanceo y calibración de umbrales, así como la exploración de algoritmos adicionales (p.ej., redes neuronales profundas) para evaluar si pueden ofrecer mejoras significativas. Asimismo, sería valioso integrar datos clínicos adicionales (como imágenes o historiales más detallados) y evaluar la generalización del modelo en cohortes de pacientes del mundo real. En conclusión, las herramientas de ML estudiadas proporcionan un enfoque prometedor para la predicción proactiva de enfermedades cardíacas, apoyando a los médicos en la identificación temprana de individuos de alto riesgo y contribuyendo potencialmente a reducir la carga de las ECV en la población.

## References


[1] World Health Organization, “Cardiovascular diseases (CVDs) Fact Sheet,” 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] V. Marina *et al.*, “Evaluating Binary Classifiers for Cardiovascular Disease Prediction,”, 2024. [Online]. Available: https://www.mdpi.com/2308-3425/11/12/396.

[3] A. S. Osei-Nkwantabisa and R. Ntumy, “Classification and Prediction of Heart Diseases using Machine Learning Algorithms,”, 2023. [Online]. Available: https://arxiv.org/pdf/2409.03697

[4] D. Dua and C. Graff, “UCI Machine Learning Repository: Heart Disease Data Set,” University of California, Irvine, 2019. [Online]. Available: https://archive.ics.uci.edu/dataset/45/heart+disease.

[5] F. Soriano, “Heart Failure Prediction Dataset,” Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction.

[6] A. Bhardwaj, “Framingham Heart Study Dataset,” Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset.