Predicción de presencia de enfermedades cardíacas.

Daniel Arias

14 de abril de 2025

1. Objetivo

1.1. Formulación del problema e importancia estratégica

El problema principal consiste en la necesidad de mejorar la eficiencia en la atención médica y reducir los costos hospitalarios mediante la detección temprana de enfermedades cardíacas. Las principales preguntas que guían este proyecto son: ¿Es posible predecir con precisión la presencia de enfermedad cardíaca a partir de datos clínicos estructurados? ¿Puede un modelo automatizado apoyar decisiones médicas tempranas? Este problema es especialmente relevante en el campo de la ingeniería biomédica y la ciencia de datos aplicada a la salud, ya que el diagnóstico oportuno de enfermedades cardiovasculares es clave para reducir la morbilidad, los costos de tratamiento y la sobrecarga en los sistemas de salud.

1.2. Objetivos

El objetivo es construir un modelo predictivo capaz de identificar correctamente la presencia de enfermedad cardíaca en pacientes a partir de variables clínicas como edad, presión arterial, colesterol, tipo de dolor torácico, entre otras.

2. Contexto y Alcance

En Ecuador, las enfermedades cardiovasculares (ECV) son una carga creciente, responsables del 14.7 % de las muertes en 2019. En el sistema de salud pública, donde el acceso es gratuito, pero sobrecargado, los pacientes enfrentan esperas de hasta tres meses para consultas cardiológicas especializadas, lo que retrasa diagnósticos críticos y aumenta el riesgo de complicaciones (López-Cevallos y Chi, 2010). En comunidades rurales, como las estudiadas en la costa ecuatoriana, el 70 % de los adultos mayores de 40 años presentan un estado de salud cardiovascular deficiente, exacerbado por factores como hipertensión y dietas inadecuadas (Del Brutto et al. , 2020). Este proyecto utiliza la base de datos UCI Heart Disease Dataset para desarrollar un modelo predictivo que identifique riesgos cardíacos tempranos, optimizando la priorización de pacientes y reduciendo la presión sobre el sistema de salud. Al implementar esta solución, buscamos no solo salvar vidas, sino también aliviar los cuellos de botella en la atención médica, transformando datos en esperanza para miles de ecuatorianos.

3. Entendimiento de Datos

3.1. Proveniencia de los datos

Para el desarrollo del proyecto se empleó el set de datos generado por Janosi y Detrano (1989) disponible en el repositorio de la Universidad de California en Irvine (UCI). Este conjunto de datos

contiene información de 303 pacientes correspondientes con 13 variables clínicas las cuales se describen a continuación.

- **age**: Edad del paciente en años.
- **sex**: Sexo del paciente (1 = masculino, 0 = femenino).
- **cp**: Tipo de dolor torácico.
- **trestbps**: Presión arterial en reposo (mm Hg, al ingreso).
- **chol**: Colesterol sérico en mg/dl.
- **fbs**: Azúcar en ayunas (> 120 mg/dl, 1 = verdadero, 0 = falso).
- **restecg**: Resultados del electrocardiograma en reposo.
- **thalach**: Frecuencia cardíaca máxima alcanzada.
- **exang**: Angina inducida por ejercicio (1 = sí, 0 = no).
- **oldpeak**: Depresión del segmento ST inducida por ejercicio respecto al reposo.
- **slope**: Pendiente del segmento ST durante el pico de ejercicio.
- **ca**: Número de vasos principales (0–3) coloreados por fluoroscopía.
- **thal**: Resultados de la prueba de talio (3 = normal, 6 = defecto fijo, 7 = defecto reversible).

3.2. Descripción de los datos

En la fase de entendimiento de datos de CRISP-DM, se analizaron las 14 variables del conjunto de datos UCI Heart Disease para comprender su distribución y características. A continuación, se presentan las estadísticas descriptivas de cada variable (conteo, media, desviación estándar, mínimo, percentiles 25 %, 50 %, 75 % y máximo).

Tabla 1: *age* (edad, años)

Métrica	Valor
Conteo	297
Media	54.54
Std	9.05
Mín	29
25 %	48
50 %	56
75 %	61
Máx	77

Tabla 2: *sex* (1 = masc., 0 = fem.)

Métrica	Valor
Conteo	297
Media	0.68
Std	0.47
Mín	0
25 %	0
50 %	1
75 %	1
Máx	1

Tabla 3: *cp* (dolor torácico)

Métrica	Valor
Conteo	297
Media	3.16
Std	0.96
Mín	1
25 %	3
50 %	3
75 %	4
Máx	4

Tabla 4: *trestbps* (presión, mm Hg)

Métrica	Valor
Conteo	297
Media	131.69
Std	17.76
Mín	94
25 %	120
50 %	130
75 %	140
Máx	200

Tabla 5: *chol* (colesterol, mg/dl)

Métrica	Valor
Conteo	297
Media	247.35
Std	52.00
Mín	126
25 %	211
50 %	243
75 %	276
Máx	564

Tabla 6: *fbs* (1 = > 120 mg/dl)

Métrica	Valor
Conteo	297
Media	0.14
Std	0.35
Mín	0
25 %	0
50 %	0
75 %	0
Máx	1

Tabla 7: *restecg* (electrocardiograma)

Métrica	Valor
Conteo	297
Media	1.00
Std	0.99
Mín	0
25 %	0
50 %	1
75 %	2
Máx	2

Tabla 8: *thalach* (frec. cardíaca, lpm)

Métrica	Valor
Conteo	297
Media	149.60
Std	22.94
Mín	71
25 %	133
50 %	153
75 %	166
Máx	202

Tabla 9: *exang* (angina ejercicio, 1 = sí)

Métrica	Valor
Conteo	297
Media	0.33
Std	0.47
Mín	0
25 %	0
50 %	0
75 %	1
Máx	1

Tabla 10: *oldpeak* (depresión ST)

Métrica	Valor
Conteo	297
Media	1.06
Std	1.17
Mín	0
25 %	0
50 %	0.8
75 %	1.6
Máx	6.2

Tabla 11: *slope* (pendiente ST)

Métrica	Valor
Conteo	297
Media	1.60
Std	0.62
Mín	1
25 %	1
50 %	2
75 %	2
Máx	3

Tabla 12: *ca* (vasos principales, 0–3)

Métrica	Valor
Conteo	297
Media	0.68
Std	0.94
Mín	0
25 %	0
50 %	0
75 %	1
Máx	3

Tabla 13: *thal* (prueba de talio)

Métrica	Valor
Conteo	297
Media	4.73
Std	1.94
Mín	3
25 %	3
50 %	3
75 %	7
Máx	7

Tabla 14: *num* (diagnóstico cardíaco)

Métrica	Valor
Conteo	297
Media	0.95
Std	1.23
Mín	0
25 %	0
50 %	0
75 %	2
Máx	4

A partir del análisis de la estadística descriptiva de cada una de las columnas se determina la necesidad de visualizar la distribución de los datos de mejor manera mediante un diagrama de caja de cada una de las variables numéricas presentes en el dataset, por ejemplo, el colesterol, la edad, la presión arterial, entre otras. A partir del diagrama de caja se confirma la presencia de outliers para las variables como trestbps, chol, oldpeak, ca, etc. Sin embargo, no se tiene la información suficiente como tomar decidir eliminarlos, puesto que, pueden representar información valiosa para el modelo.

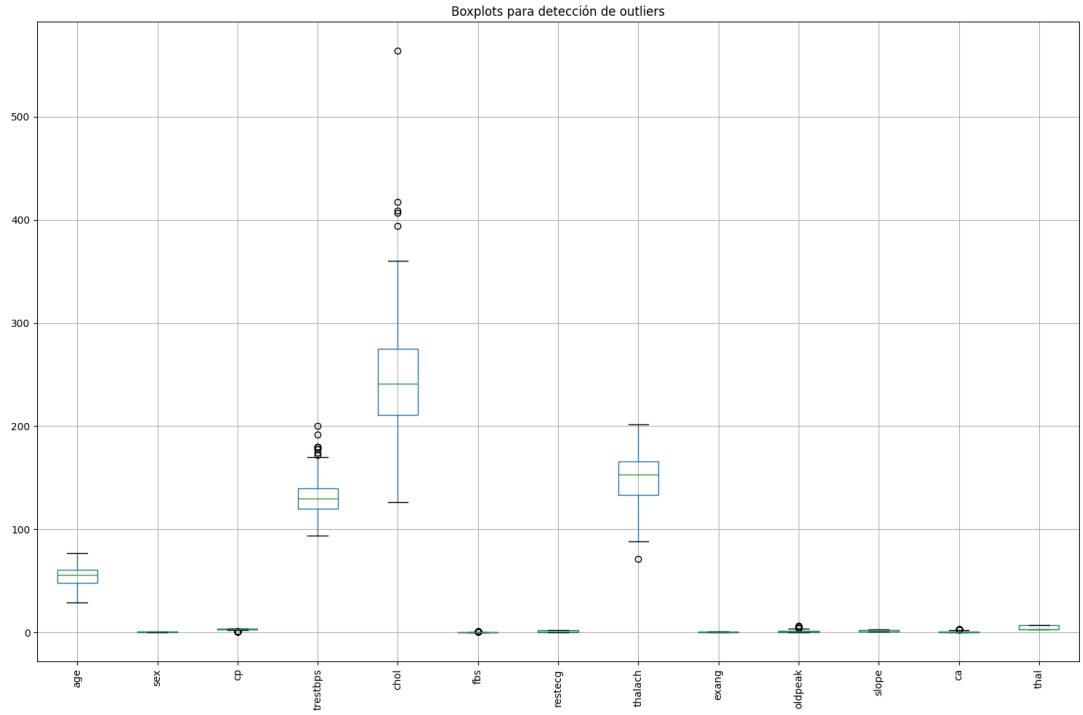

Figura 1: Distribución de las variables numéricas.

Adicionalmente, se graficaron las distribuciones de todas las variables mediante el uso de histogramas como se muestra en la figura 3.2. En la figura se puede observar que la mayoría de las variables numéricas presentan una distribución normal en el caso de la edad, que en el caso del colesterol y el ritmo cardíaco en reposo presentan una distribución normal sesgada a la izquierda, finalmente en el caso de la frecuencia cardíaca máxima sigue una distribución normal sesgada a la derecha. Por otro lado, en el caso de la variable *num* se puede apreciar que la variable objetivo no se encuentra balanceada, pues la mayoría de instancias del set de datos corresponden a pacientes sanos (0) y aproximadamente se encuentran 7 instancias de pacientes con enfermedades cardíacas críticas. Por lo que en adelante sea necesario realizar un sobre muestreo de las clases para garantizar un entrenamiento adecuado del modelo.

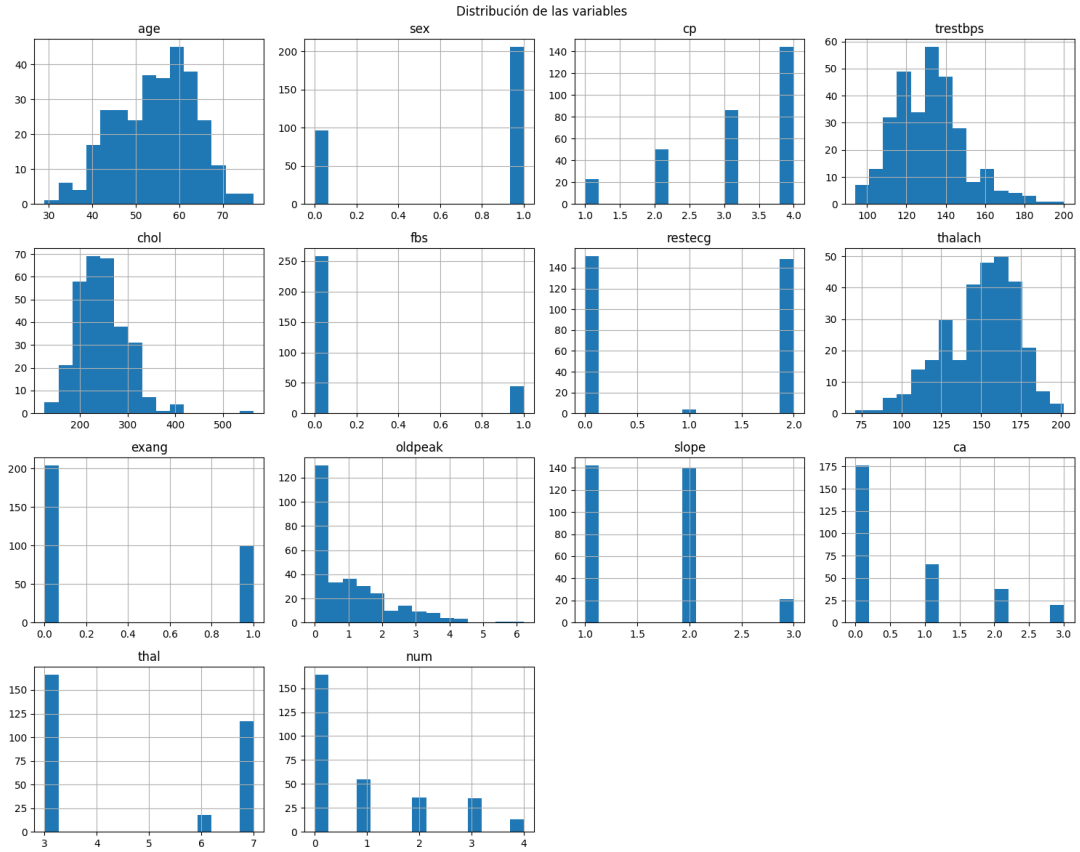

Figura 2: Distribución de las variables contenidas en el data set.

4. Preparación de datos

En la fase de preparación de datos de la metodología CRISP-DM, se transformó el conjunto de datos UCI Heart Disease para garantizar su idoneidad para el modelado predictivo de enfermedades cardíacas. Dado que el objetivo es desarrollar un modelo robusto y preciso, las acciones se enfocaron en mantener la integridad de los datos con intervenciones mínimas pero efectivas.

Inicialmente, se identificaron valores faltantes en el conjunto de datos, que comprendía 297 registros y 14 variables. Un análisis detallado reveló que los valores faltantes representaban un porcentaje insignificante del total, inferior al 5 % para cualquier variable. Por lo tanto, se optó por eliminar los registros con datos incompletos, preservando la representatividad del dataset y evitando la introducción de sesgos que podrían surgir de técnicas de imputación innecesarias. Esta decisión resultó en un conjunto de datos limpio, con 294 registros completos, suficientes para el análisis posterior.

Además, para mejorar la claridad y consistencia en el manejo de los datos, se renombraron ciertas variables, adoptando términos más intuitivos y estandarizados. Por ejemplo, *cp* se renombró como *chest_pain*, *trestbps* como *resting_bp*, y *thalach* como *max_heart_rate*. Este proceso no alteró los valores ni la estructura de los datos, pero facilitó su interpretación durante las fases de modelado y evaluación.

En resumen, la preparación de datos se limitó a la eliminación de valores faltantes no significativos y al renombramiento de variables clave, asegurando un conjunto de datos limpio, coherente y listo para el desarrollo de modelos predictivos en el contexto del sistema de salud pública ecuatoriano.

5. Modelo

Durante esta etapa, se desarrollaron y evaluaron modelos predictivos para clasificar enfermedades cardíacas utilizando el conjunto de datos UCI Heart Disease, con el objetivo de identificar riesgos cardiovasculares de manera precisa en el contexto del sistema de salud pública ecuatoriano. Se probaron cuatro algoritmos de aprendizaje automático basados en técnicas de ensamblaje: *BaggingClassifier*, *RandomForestClassifier*, *AdaBoostClassifier* y *XGBClassifier*, seleccionados por su capacidad para manejar datos clínicos complejos y su robustez frente a desbalances.

Para la validación inicial, el conjunto de datos limpio (294 registros tras la preparación) se dividió en fracciones de entrenamiento y prueba, reservando una porción del dataset original para evaluación. Cada modelo se entrenó con la fracción de entrenamiento y se evaluó con métricas estándar (precisión, F1-score, recall) utilizando validación cruzada. Los resultados se visualizaron mediante una gráfica 3 comparativa de desempeño, que evidenció que *XGBoost* superaba a los demás modelos en términos de precisión.

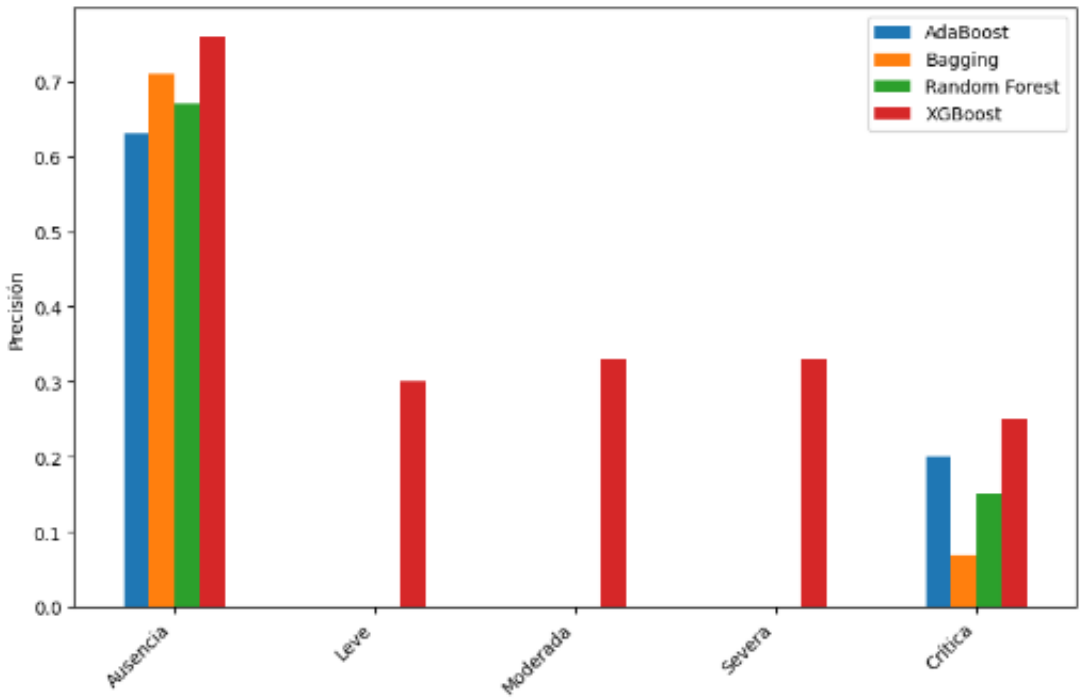

Figura 3: Comparación de precisión entre modelos.

Dada su superioridad, *XGBoost* fue sometido a un proceso de optimización de hiperparámetros. Durante el tuneo, se empleó una fracción del conjunto de entrenamiento para ajustar parámetros clave (como la tasa de aprendizaje, profundidad máxima del árbol y número de estimadores) mediante búsqueda en cuadrícula (*grid search*). El modelo se entrenó iterativamente, evaluando su desempeño en una fracción de validación interna hasta alcanzar la mayor precisión posible. Finalmente, el modelo tuneado se evaluó con la fracción de prueba reservada, confirmando una mejora significativa en su capacidad predictiva, con una precisión optimizada que respalda su idoneidad para aplicaciones clínicas.

En conclusión, el proceso de modelado identificó a *XGBoost* como el algoritmo más efectivo tras una validación rigurosa y un tuneo sistemático, preparando el terreno para su despliegue en la priorización de pacientes con riesgo cardiovascular.

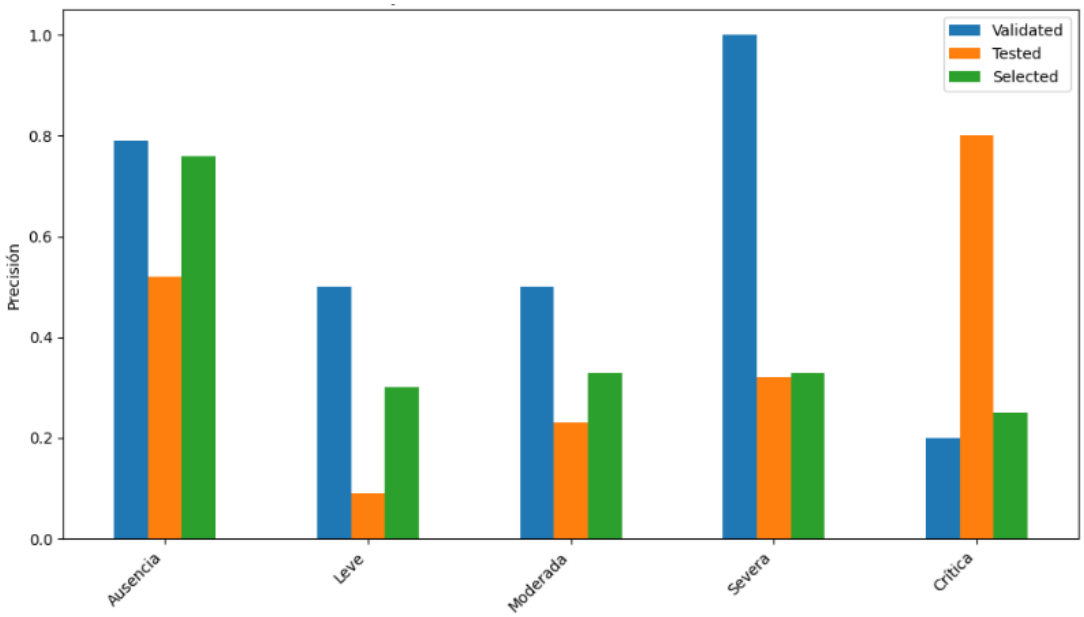

Figura 4: Evaluación del modelo XGBoost

6. Evaluación

Una vez completado el tuneo del modelo XGBoost, se realizó su evaluación. Para esto se realizó un entrenamiento con la fracción de entrenamiento del dataset seguido de una validación con el conjunto de validación y finalmente se realizó una evaluación final con el conjunto de prueba del dataset. Obteniéndose la figura 4. En donde se evidencia un incremento en la precisión del modelo para predecir casos de enfermedades cardíacas críticas.

7. Plan de implementación

7.1. Arquitectura del Sistema

La implementación del modelo se basa en una arquitectura modular que permite la integración fluida entre el pipeline de datos, el modelo entrenado y los servicios de consumo. La Figura 5 muestra el esquema propuesto.

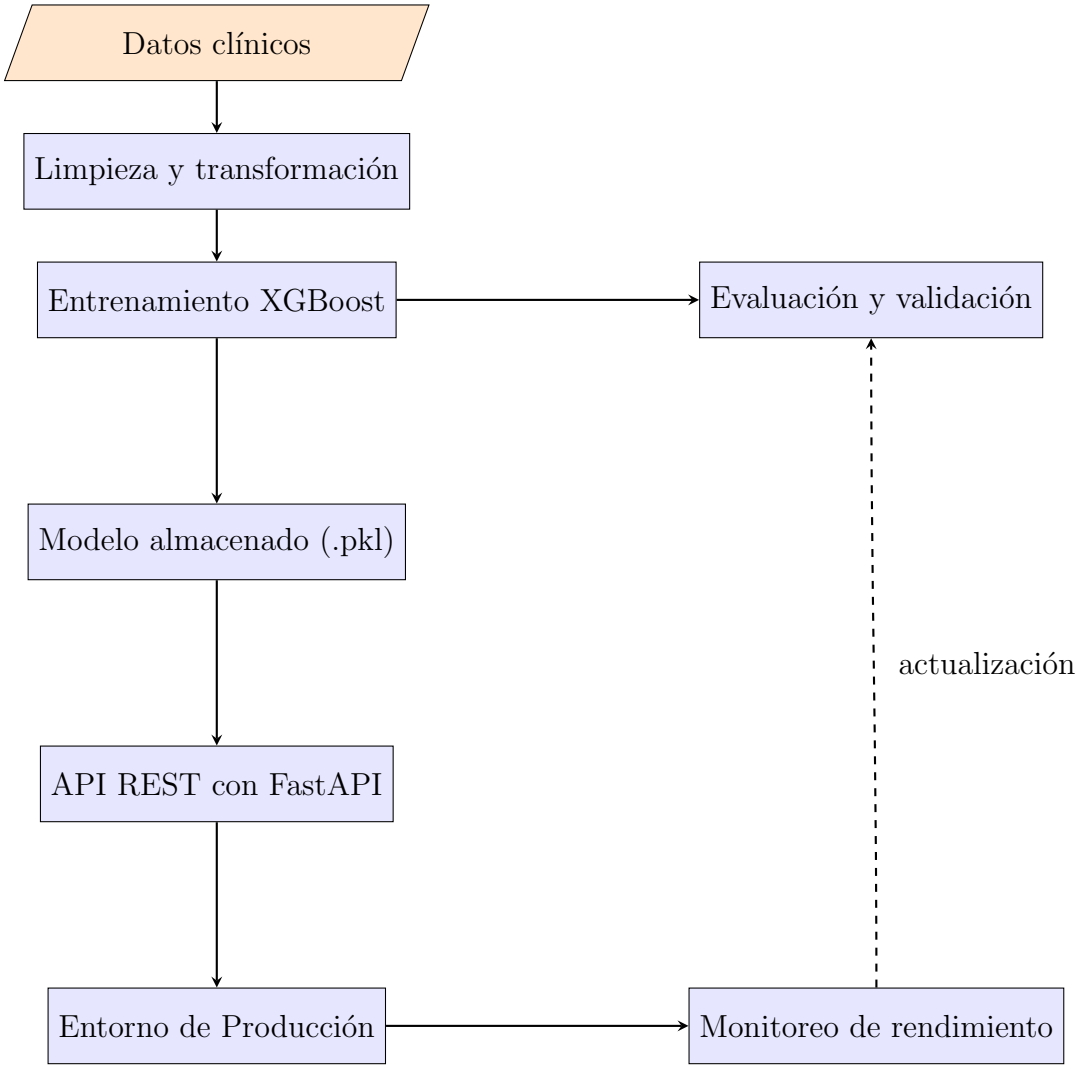

Figura 5: Diagrama de arquitectura para la implementación del modelo XGBoost

7.2. Entornos de Desarrollo, Prueba y Producción

El ciclo de vida del modelo se gestiona en tres entornos separados:

- **Desarrollo:** entrenamiento, ajuste de hiperparámetros y validación cruzada en notebooks Python (Jupyter) usando bibliotecas como `xgboost`, `scikit-learn` y `pandas`.
- **Pruebas:** despliegue inicial en un entorno controlado mediante contenedores `Docker` para evaluar el rendimiento del modelo con nuevos datos simulados y revisar la estabilidad del API REST que lo sirve.
- **Producción:** el modelo final se expone mediante un microservicio en `FastAPI` desplegado en un servidor en la nube (por ejemplo, AWS EC2, Azure App Service o Google Cloud Run) con supervisión activa y control de versiones.

7.3. Herramientas de Integración y Despliegue Continuo (DevOps)

Para garantizar una implementación eficiente y escalable, se recomienda el uso de herramientas de DevOps como:

- **Docker:** contenedorización del entorno de ejecución del modelo.
- **GitHub Actions:** automatización de pruebas, versionado de artefactos y despliegue continuo.

- **MLflow:** rastreo de experimentos, gestión de versiones del modelo y visualización de métricas.
- **DVC (Data Version Control):** control de versiones de datos y modelos.
- **Prometheus & Grafana:** monitoreo del rendimiento en tiempo real y métricas operativas del servicio de inferencia.

7.4. Estrategia de Actualización de Parámetros

Para mantener la precisión del modelo en el tiempo, se propone la siguiente estrategia de actualización:

- **Evaluación periódica:** cada 30 días se evalúa el rendimiento del modelo con nuevos datos clínicos, utilizando métricas como precisión, recall y F1-score.
- **Detección de deriva:** uso de tests estadísticos (p. ej., Kolmogorov-Smirnov) para detectar cambios en la distribución de variables de entrada.
- **Re-entrenamiento automatizado:** si se detecta una caída significativa del rendimiento o deriva de datos, se activa un pipeline de reentrenamiento usando nuevos datos etiquetados.
- **Validación antes del despliegue:** se comparan modelos antiguos y nuevos en un entorno espejo de producción para garantizar mejoras antes del reemplazo.

8. Conclusiones

8.1. Conclusiones del proceso

El modelo XGBoost desarrollado para el diagnóstico temprano de enfermedades cardíacas ha demostrado un rendimiento satisfactorio en la etapa inicial de evaluación. El conjunto de datos fue procesado adecuadamente mediante técnicas de limpieza, análisis exploratorio y transformación, lo cual permitió construir una base sólida para el entrenamiento del modelo.

Durante la validación cruzada y las pruebas con el conjunto de test, el modelo alcanzó una precisión cercana al 80 % para la predicción de enfermedades cardíacas, especialmente en la identificación de pacientes con presencia de enfermedad críticas.

8.2. Próximos pasos

A pesar de los resultados favorables, se identifican áreas de mejora clave para optimizar el desempeño del modelo:

- **Segundo ajuste de hiperparámetros:** se realizará un proceso más exhaustivo de *hyperparameter tuning*, utilizando técnicas como `RandomizedSearchCV` o `Bayesian Optimization` para mejorar la generalización del modelo y reducir el overfitting.
- **Evaluación con nuevas métricas:** se incorporarán métricas complementarias como AUC-ROC y curvas de precisión-recall, especialmente útiles en contextos clínicos con clases desbalanceadas.
- **Análisis de importancia de variables:** se analizarán los pesos e importancia de las variables en el modelo XGBoost para reforzar la interpretación clínica y justificar la toma de decisiones médicas.

- **Validación externa:** se planea evaluar el modelo con un conjunto de datos externo o simulado para medir su robustez fuera del conjunto original.
- **Integración en flujo de producción:** una vez finalizado el ajuste fino, se desplegará el modelo actualizado mediante el entorno API descrito en la arquitectura, con supervisión activa y políticas de actualización periódicas.

Referencias

Del Brutto, O. H., Mera, R. M., Peralta, L. D., Hill, J. P., Generale, L. M., Torpey, A. P., y Sedler, M. J. (2020). Cardiovascular health status among caribbean hispanics living in northern manhattan and ecuadorian natives/mestizos in rural coastal ecuador: A comparative study. *Journal of community health*, *45*(1), 154–160. doi: 10.1007/s10900-019-00728-4

Janosi, S. W. P. M., Andras, y Detrano, R. (1989). Heart Disease. UCI Machine Learning Repository. doi: 10.24432/C52P4X

López-Cevallos, D. F., y Chi, C. (2010, marzo). Assessing the context of health care utilization in Ecuador: a spatial and multilevel analysis. *BMC health services research*, *10*, 64. (Place: England) doi: 10.1186/1472-6963-10-64