Parcial Aprendizaje Automático

Profesor: Nicolas, Caballero.

Estudiante: Guillermo, Carcamo.

Diagnóstico de Endometriosis mediante Modelos de Aprendizaje Automático

1. Introducción

La endometriosis es una enfermedad ginecológica crónica caracterizada por el crecimiento anormal de tejido endometrial fuera del útero.
Afecta a un porcentaje considerable de mujeres en edad fértil y su diagnóstico suele ser tardío debido a la complejidad clínica y a la inespecificidad de los síntomas.

El presente trabajo busca aplicar técnicas de **aprendizaje automático supervisado** para predecir el diagnóstico de endometriosis a partir de variables clínicas simuladas.
El proyecto integra todas las etapas del flujo de trabajo de *Machine Learning*, desde la exploración de datos hasta la comparación de modelos y la interpretación de resultados.

2. Descripción del Dataset

El conjunto de datos utilizado contiene **10.000 registros sintéticos** provenientes de una fuente pública de Kaggle.
Los datos fueron generados artificialmente con fines educativos y no representan información clínica real.

Variables principales:

- **Age:** Edad de la paciente.
- **BMI:** Índice de masa corporal.
- **Chronic_Pain_Level:** Nivel de dolor crónico (escala 1–10).
- **Menstrual_Irregularity:** Presencia de irregularidades menstruales (0/1).
- **Hormone_Level_Abnormality:** Anormalidades hormonales (0/1).
- **Infertility:** Dificultad para concebir (0/1).
- **Diagnosis:** Variable objetivo (0 = No Endometriosis, 1 = Endometriosis).

Los datos fueron analizados de forma exploratoria para conocer su estructura, distribución y posibles relaciones entre variables.

3. Análisis Exploratorio de Datos

Durante la fase de exploración se realizaron análisis estadísticos y visuales con el fin de comprender la composición del dataset.

3.1 Distribución y correlaciones

Se observaron edades comprendidas entre 20 y 50 años, con una distribución relativamente uniforme, evitando sesgos por rango etario.
El mapa de calor de correlaciones mostró **baja colinealidad** entre las variables numéricas, lo cual permite entrenar modelos sin riesgo de redundancia informativa.

3.2 Relación entre variables clínicas y diagnóstico

El análisis de frecuencias mostró una mayor proporción de anormalidades hormonales en los casos positivos de endometriosis, lo que sugiere una relación clínica razonable.
Sin embargo, las correlaciones directas fueron moderadas, lo que motivó la exploración de modelos más complejos.

3.3 PCA exploratorio

Se aplicó un **Análisis de Componentes Principales (PCA)** con fines exploratorios, para analizar la estructura interna del dataset.
Las dos primeras componentes no mostraron una separación clara entre clases, evidenciando que el diagnóstico depende de **combinaciones no lineales de variables**.
La primera componente estuvo más influenciada por **irregularidad menstrual** e **infertilidad**, y la segunda por **IMC** y **nivel de dolor crónico**.

4. Metodología y Preprocesamiento

4.1 División de datos

El dataset se dividió en **70% para entrenamiento** y **30% para prueba** utilizando la función *train_test_split* con estratificación, garantizando la proporción de clases en ambos conjuntos.
Esta división permite evaluar la capacidad de generalización de los modelos en datos no vistos.

4.2 Escalado de variables

Se aplicó **StandardScaler** a las variables numéricas, ajustando cada una a una media de 0 y desviación estándar de 1.
Esto es fundamental en modelos sensibles a la escala, como la Regresión Logística, para evitar que variables con valores más altos dominen el aprendizaje.

4.3 Modelos utilizados

Se entrenaron tres modelos de clasificación:

1. **Regresión Logística:** modelo lineal probabilístico.
2. **Árbol de Decisión:** modelo no lineal basado en divisiones sucesivas.
3. **Random Forest:** modelo de ensamble que combina múltiples árboles para mejorar la estabilidad y precisión.

Cada modelo fue evaluado mediante métricas de desempeño y comparado para determinar el mejor rendimiento global.

5. Resultados y Evaluación

5.1 Regresión Logística

- **Accuracy:** 0.625
- **Precision:** 0.562
- **Recall:** 0.364
- **F1-score:** 0.441
- **ROC-AUC:** 0.654

El modelo distingue moderadamente bien entre clases, aunque presenta un **recall bajo**, lo que implica que detecta solo el 36% de los casos positivos.
La matriz de confusión muestra 1429 verdaderos negativos, 445 verdaderos positivos y 779 falsos negativos.
Se trata de un modelo **conservador**, con buena precisión pero baja sensibilidad.

5.2 Árbol de Decisión

- **Accuracy:** 0.539
- **Precision:** 0.438
- **Recall:** 0.462
- **F1-score:** 0.450
- **ROC-AUC:** 0.527

El Árbol de Decisión logra detectar más casos positivos que la Regresión Logística (recall más alto), pero comete más errores globales y tiene menor capacidad discriminante (ROC-AUC menor).
Es un modelo **más sensible**, aunque menos preciso.

5.3 Comparativa de modelos base

Métrica	Regresión Logística	Árbol de Decisión	Mejor
Accuracy	0.625	0.539	Regresión Logística
Precision	0.562	0.438	Regresión Logística
Recall	0.364	0.462	Árbol de Decisión
F1-score	0.441	0.450	Empate
ROC-AUC	0.654	0.527	Regresión Logística

Ambos modelos mostraron un rendimiento moderado, con comportamientos complementarios:
la **Regresión Logística** fue más precisa, y el **Árbol de Decisión** más sensible.

5.4 Random Forest (modelo avanzado)

El modelo de **Random Forest** permitió mejorar la estabilidad y la interpretación de las variables.
Las más influyentes fueron:

- **Chronic_Pain_Level (dolor crónico)**
- **BMI (índice de masa corporal)**
- **Age (edad)**

Esto demuestra que el dolor persistente y el estado físico general tienen un papel clave en el diagnóstico, mientras que las variables hormonales y menstruales aportan menor peso predictivo.
El modelo mejoró el equilibrio entre precisión y recall, mostrando un desempeño más consistente que los anteriores.

6. Discusión

Los resultados reflejan las limitaciones y ventajas de cada enfoque.
La Regresión Logística ofreció un **rendimiento más estable** y una mejor discriminación global (AUC = 0.65),
mientras que el Árbol de Decisión presentó un **mayor recall**, útil en contextos donde se prioriza la detección de casos positivos.

El Random Forest combinó ambas fortalezas, logrando un mejor equilibrio y permitiendo interpretar la **importancia de las variables**.
El uso de datos sintéticos explica la moderada correlación entre síntomas clínicos y diagnóstico,
pero aun así los modelos lograron identificar patrones coherentes.

7. Conclusiones

El proyecto permitió integrar todo el proceso de aprendizaje automático:
desde la exploración de datos, preprocesamiento y modelado, hasta la evaluación comparativa.

Los modelos lograron resultados consistentes, destacando que el **Random Forest** fue el más informativo,
identificando como variables más relevantes el **dolor crónico**, el **IMC** y la **edad**.

Más allá de las métricas, el principal aprendizaje fue comprender cómo cada algoritmo interpreta los datos de manera distinta y cómo las decisiones de preprocesamiento (como el escalado y la división 70/30) impactan en el resultado final.