Universidad de Guadalajara

Licenciatura en *Inteligencia Artificial y Ciencia de Datos*

Proyecto Integrador Matemáticas Aplicadas

Matemáticas Aplicadas a la Ciencia de Datos

Impartida por

Mtro. IVAN ALEJANDRO TOLEDANO JUAREZ

P r e s e n t a

Iris Alina Pérez Rivera

Fernanda Garcia Rodriguez

Diana García Trujillo

Karol Paola Rosales Miranda

Zapopan, Jalisco; 29 de Noviembre de 2025

Introducción

El Síndrome de Ovario Poliquístico (PCOS) es un trastorno endocrino con implicaciones reproductivas y metabólicas. En este trabajo se evalúan varios modelos de clasificación supervisada sobre un dataset clínico (541 pacientes, ~44 variables) con el propósito de identificar un modelo adecuado para detectar pacientes con PCOS. Tras comparar Random Forest, Árbol de Decisión, SVM y Regresión Logística, el mejor balance entre interpretabilidad y desempeño fue Regresión Logística. Este documento presenta la justificación matemática del modelo seleccionado y su aplicación práctica al dataset.

Planteamiento del problema

Clasificación binaria para predecir si una paciente tiene PCOS (1) o no (0) a partir de variables demográficas, clínicas y bioquímicas.
Nuestro objetivo es seleccionar y justificar matemáticamente un modelo y demostrar su aplicación práctica sobre el dataset PCOS_data_1.xlsx.

Descripción del dataset

Nuestro dataset es un archivo clínico anonimizado "PCOS_data_1.xlsx". Cuenta con 541 pacientes. Con 44 variables: (edad, peso, BMI, recuento de folículos derecho/izquierdo, niveles hormonales FSH, LH, TSH, AMH, signos clínicos como hair growth, skin darkening, ciclo menstrual, etc.). Se observó un desbalance de aproximadamente 67% sin PCOS (0) y 33% con PCOS (1).

Nuestro preprocesamiento realizado se basó en la normalización de nombres de columna; imputación (mediana para numéricas, moda para binarias); eliminación de columnas no relevantes; estandarización para algoritmos sensibles (StandardScaler); selección de características basada en correlación y eliminación de multicolinealidad (|r|>0.8).

Desarrollo teórico

Formulación matemática del modelo

La regresión logística modela la probabilidad condicional de la clase 1 (PCOS) como:

$$P(y=1 \mid x) = \sigma(z) = \frac{1}{1+e^{-z}}, \quad z = w^T x + b,$$

donde x ∈ R^d es el vector de características, w ∈ R^d son los coeficientes y b el sesgo (intercepto). Para clasificación binaria, la regla de decisión típica es:

$$\hat{y} = \begin{cases} 1 & \text{si } P(y=1 \mid x) \geq 0.5, \\ 0 & \text{si } P(y=1 \mid x) < 0.5. \end{cases}$$

Función de costo (log-loss / cross-entropy)

La función objetivo que se minimiza para N observaciones es la pérdida logarítmica (cross-entropy):

$$\mathcal{L}(w,b) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log p_i + (1-y_i)\log(1-p_i)\right],$$

donde $p_i = \sigma(w^T x_i + b)$. Con regularización L2 (penalización Ridge) la función se extiende a:

$$\mathcal{L}_{reg}(w,b) = \mathcal{L}(w,b) + \frac{\lambda}{2}\|w\|_2^2.$$

Algoritmo de optimización

La función de pérdida sin regularización es convexa en (w,b). Por tanto, cualquier mínimo local es mínimo global.

Algoritmos usados en la práctica: métodos de gradiente (Gradient Descent / Stochastic Gradient Descent), métodos quasi-Newton como L-BFGS o solucionadores especializados como liblinear. Otra alternativa clásica es Newton-Raphson / IRLS (Iteratively Reweighted Least Squares), que aprovecha la forma exponencial de la verosimilitud para converger rápido cuando el conjunto no es demasiado grande.

En scikit-learn, los solvers típicos: 'liblinear' (bueno para datasets más pequeños y L1), 'lbfgs' (L-BFGS, para L2 y problemas de mayor dimensión), etc. En el pipeline del notebook se evaluaron varios solvers y valores de C mediante GridSearchCV (optimización sobre F1).

Esto funciona porque la función de pérdida es diferenciable y convexa; los métodos iterativos aprovechan el gradiente para encontrar el mínimo global de manera estable.

Propiedades matemáticas relevantes

La convexidad garantiza existencia y unicidad del minimizador (con penalización L2).

Interpretabilidad de los coeficientes w se interpretan (en log-odds), un coeficiente positivo aumenta la probabilidad logarítmica de la clase 1.

Linealidad en el espacio de características de la frontera de decisión es lineal en x (aunque transformaciones de características permiten no linealidad).

Supuestos implícitos de la independencia condicional de observaciones, relación lineal entre características y log-odds (si hay interacciones o no linealidad fuerte, puede degradarse el desempeño).

Sensibilidad a multicolinealidad donde los coeficientes pueden volverse inestables si las variables están fuertemente correlacionadas.

Robustez a outliers, no es tan robusta a outliers como algunos métodos robustos; escalado y detección de outliers ayudan.

Argumento teórico para este problema

Datos clínicos con variables interpretable y un objetivo binario ya que la Regresión Logística entrega probabilidades directamente y coeficientes clínicamente interpretables, por ejemplo el efecto del recuento de folículos sobre la probabilidad log-odds de PCOS.

Tamaño de muestra moderado y número de variables, la regularización y selección de características hacen a la logística adecuada gracias al equilibrio entre complejidad e interpretabilidad.

En un escenario clínico es preferible un modelo interpretable con alta sensibilidad para minimizar falsos negativos; la logística puede ajustarse para privilegios recall mediante thresholds o weights en la función de costo.

Aplicación práctica del modelo

La aplicación práctica sobre el dataset PCOS_data_1.xlsx comenzó con la lectura y limpieza de los 541 registros y ~44 variables, normalizando nombres de columnas, realizando imputación de valores faltantes (mediana para variables continuas y moda para binarias) y eliminando columnas no informativas; a continuación se estandarizaron las características y se aplicó una selección de variables basada en correlación usando únicamente el conjunto de entrenamiento, eliminando pares con multicolinealidad ($|r|>0.8$) para conservar predictoras no redundantes.

Los datos se dividieron de forma estratificada en entrenamiento y prueba (80/20, random_state=42) para mantener la proporción de clases; el modelo se construyó mediante un pipeline que incluía StandardScaler seguido de LogisticRegession y se optimizaron hiperparámetros (C, solver y class_weight) con GridSearchCV usando validación estratificada de 5 folds y la métrica F1 como objetivo.

El mejor estimador se reentrenó con todo el conjunto de entrenamiento y se evaluó en el conjunto de prueba. En la validación cruzada el F1 promedio en validación fue cercano a 0.8393, mostrando estabilidad en las métricas; en la evaluación final sobre test el modelo obtuvo Accuracy = 0.9083, Precision = 0.8421, Recall = 0.8889, F1 = 0.8649 y ROC-AUC = 0.9589, con matriz de confusión reportada de 70 verdaderos negativos, 3 falsos positivos, 4 falsos negativos y 32 verdaderos positivos; además se extrajeron y analizaron los coeficientes del modelo final para interpretar la influencia de las variables, identificando consistentemente como más influyentes Follicle No. (R), Follicle No. (L), hair growth, skin darkening y weight gain.

Discusión

La conformidad entre las variables más importantes en modelos lineales y en métodos no lineales refuerza la validez clínica de los predictores.

La regresión logística alcanza un excelente AUC-ROC y un recall alto, lo cual es deseable en un contexto clínico donde priorizar identificación de casos positivos es crítico. Añadiendo que la interpretabilidad de la regresión logística facilita la adopción clínica donde los coeficientes permiten discutir cómo cada variable incrementa o reduce el riesgo.

Podría haber limitaciones si existieran interacciones o no linealidad no modelada, la logística puede subestimar relaciones complejas, pero la evidencia empírica aquí sugiere que las relaciones lineales en log-odds son suficientes para este dataset. Además, el desbalance moderado fue tratado con class_weight y selección de métricas (F1) en la búsqueda.

Conclusiones

El modelo Regresión Logística fue seleccionado como el mejor por su mejor balance entre rendimiento (F1 y AUC), interpretabilidad y eficiencia computacional.Mostrando resultados empíricos con: Accuracy ≈ 90.8%, F1 ≈ 86.5%, ROC-AUC ≈ 95.9%, Recall ≈ 88.9% (valores reportados en el notebook).

Con este modelo se recomienda desplegarlo como herramienta de apoyo clínico para triage; monitorear desempeño en datos nuevos y recalibrar el threshold de decisión para priorizar recall si se busca reducir falsos negativos; considerar añadir términos polinomiales o interacciones sólo si hay evidencia de no linealidad residual.

Referencias

- Fuente de datos: PCOS_data_1.xlsx.
- Notebook: Pérez Rivera, I. A., García Rodríguez, F., García Trujillo, D., & Rosales Miranda, K. P. (2025). ML final: Clasificación de PCOS (Jupyter notebook). Google Colab. https://colab.research.google.com/drive/1YazNngC0VY3TwnvmVSWAVlRyYdq-iuKo
- Scikit-learn documentation: LogisticRegression, GridSearchCV, StratifiedKFold — https://scikit-learn.org/