Universidad de Guadalajara

Licenciatura en *Inteligencia Artificial y Ciencia de Datos*

Proyecto Integrador Matemáticas Aplicadas a Ciencia de Datos

Matemáticas Aplicadas

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; 26 de Noviembre de 2025

Análisis Matemático de Regresión Logística para Diagnóstico de PCOS

Introducción

Contexto del Problema

El Síndrome de Ovario Poliquístico (PCOS) es una condición endocrina que afecta aproximadamente al 10% de las mujeres en edad reproductiva a nivel mundial. Se caracteriza por desequilibrios hormonales, quistes ováricos y diversos síntomas metabólicos. El diagnóstico temprano es crucial para prevenir complicaciones a largo plazo como diabetes tipo 2, enfermedades cardiovasculares e infertilidad.

Justificación

La implementación de modelos predictivos para el diagnóstico de PCOS resulta fundamental para complementar la labor médica, agilizando el proceso de detección mediante el análisis automatizado de variables clínicas y hormonales. Así mismo, estos modelos no solo asisten a los profesionales de la salud al identificar patrones complejos y relaciones no evidentes en los datos, sino que también reducen significativamente los tiempos de diagnóstico, permitiendo intervenciones tempranas.

De igual manera, pueden complementar la educación médica, al ilustrar la aplicación de tecnologías innovadoras en escenarios clínicos reales y fomentar un enfoque basado en datos para la toma de decisiones.

Descripción del Dataset

El dataset “PCOS_data” comprende información clínica de 541 pacientes, caracterizado por 44 variables que abarcan dimensiones demográficas, bioquímicas, sintomatológicas y de estilo de vida.

Entre los datos recopilados se incluyen variables antropométricas (edad, peso, altura, BMI y medidas corporales), niveles hormonales y bioquímicos (FSH, LH, TSH, AMH, prolactina, vitamina D3, glucosa y hemoglobina), indicadores clínicos y sintomatológicos (regularidad y duración del ciclo menstrual, historial reproductivo, y manifestaciones como crecimiento de vello, acné y pérdida de cabello), así como factores relacionados con el estilo de vida (hábitos alimenticios, práctica de ejercicio y presión arterial), lo que permite la identificación de patrones asociados al PCOS.

Planteamiento del Problema

El presente estudio aborda un problema de clasificación binaria en el ámbito del diagnóstico médico asistido por machine learning. La variable objetivo se define mediante dos categorías mutuamente excluyentes: la Clase 0, que corresponde a pacientes sin Síndrome de Ovario Poliquístico (PCOS), representando 364 casos que equivalen al 67.3% de la muestra; y la Clase 1, que identifica a pacientes con diagnóstico de PCOS, constituyendo 177 casos que representan el 32.7% del dataset. Esta distribución refleja la prevalencia natural de la condición en la población de estudio y establece el marco para el desarrollo de un modelo predictivo que distinga efectivamente entre ambos grupos.

Metodología y limpieza de datos

Análisis exploratorio y visualizaciones importantes

Preprocesamiento de datos

Se normalizaron los nombres de las columnas eliminando espacios y tabulaciones inconsistentes. Se detectaron 3 columnas con valores faltantes; para su imputación se usó la moda en variables categóricas binarias y la mediana en variables numéricas continuas. Estas decisiones se tomaron para minimizar el sesgo por valores extremos y mantener la interpretabilidad básica de las variables.

El preprocesamiento fue fundamental para garantizar la calidad de los datos antes del modelado. Realizamos una limpieza exhaustiva que incluyó la normalización de nombres de columnas, eliminando espacios y caracteres especiales. Para manejar los valores faltantes, implementamos una estrategia de imputación inteligente: utilizamos la mediana para variables numéricas continuas y la moda para variables binarias. También eliminamos columnas no numéricas que no aportaban al análisis, como los grupos sanguíneos. Finalmente, aplicamos estandarización a las características para que todas tuvieran la misma escala, lo que es crucial para algoritmos sensibles como SVM.

Análisis Exploratorio

El análisis exploratorio nos permitió entender la estructura y distribución de nuestros datos. Comenzamos examinando las dimensiones del dataset, identificando 541 pacientes con 42 variables clínicas cada uno. Analizamos la distribución de la variable objetivo PCOS (Y/N), encontrando un desbalance natural con 67% de pacientes sin PCOS y 33% con PCOS. Generamos visualizaciones como boxplots para comparar distribuciones de edad y BMI entre grupos, y heatmaps de correlación para identificar relaciones entre variables. Este análisis exploratorio se usó en todos los algoritmos.

Pasos realizados en análisis exploratorio

Primero, realizamos un análisis univariado examinando distribuciones individuales de cada variable. Luego, pasamos al análisis bivariado, estudiando las relaciones entre pares de variables y con la variable objetivo. Utilizamos matrices de correlación para identificar multicolinealidad entre predictores. Generamos visualizaciones comparativas como gráficos de barras para variables categóricas y boxplots para variables continuas (todas se

encuentran en el código). Finalmente, realizamos análisis de valores atípicos para detectar observaciones extremas que podrían afectar el modelado.

En el proyecto integrador de Machine Learning, identificamos que la Regresión Logística era el mejor modelo para clasificar pacientes con Síndrome de Ovario Poliquístico (PCOS), alcanzando un 93.58% de accuracy y AUC-ROC de 0.9619. Este documento profundiza en las bases matemáticas que sustentan este modelo, conectando la teoría con su aplicación práctica en el diagnóstico médico.

Descripción del Dataset

El conjunto de datos utilizado en esta investigación incluye diversas variables clínicas y demográficas recolectadas de 541 pacientes. Entre las características identificadas como más relevantes para el modelo predictivo se encuentran: el número de folículos ováricos en los ovarios izquierdo y derecho (Follicle No. R/L), que constituye un marcador ecográfico fundamental; la presencia de oscurecimiento cutáneo (Skin darkening), asociado a desequilibrios hormonales; el crecimiento excesivo de vello (Hair growth), indicativo de hiperandrogenismo; el aumento de peso (Weight gain), relacionado con alteraciones metabólicas; y la irregularidad del ciclo menstrual (Cycle regularity), considerada uno de los criterios diagnósticos centrales para el PCOS. Estas variables demostraron tener el mayor poder discriminativo según el análisis de importancia de características realizado durante la fase de modelado.

Desarrollo Teórico - Regresión Logística

Formulación Matemática del Modelo

La regresión logística modela la probabilidad de pertenecer a la clase 1 (PCOS) mediante la función sigmoide:

$$P(Y = 1|X) = \frac{1}{1 + e^{-z}}$$

donde:

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Parámetros del modelo:

- $\beta 0$: Intercepto (en nuestro modelo: -0.7386)

- $\beta 1, \beta 2, ..., \beta n$: Coeficientes de las variables predictoras
- $X1, X2, ..., Xn$: Variables clínicas (Follicle No., Hair growth, etc.)

Supuestos Matemáticos del Modelo

Para que la regresión logística funcione correctamente, necesita cumplir con ciertos supuestos matemáticos:

Primero, asume que existe una relación lineal entre las características que analizamos y lo que llamamos el "log-odds" (o logaritmo de las probabilidades) del resultado. Esto significa que si una variable aumenta, su efecto en la probabilidad de tener PCOS crece de manera constante y proporcional.

Otro supuesto importante es que cada paciente representa una observación independiente. Es decir, la información de una persona no afecta a la de otra, y todas las muestras se obtuvieron de manera representativa.

Por último, el modelo funciona mejor cuando las variables que usamos no están altamente correlacionadas entre sí. Si dos variables miden prácticamente lo mismo, puede causar inestabilidad en los resultados y dificultar entender cuál es realmente su influencia individual en el diagnóstico.

Función de Costo - Log-Loss

La función objetivo a minimizar es la entropía cruzada binaria.

Log-loss es como un parámetro en el cual si el modelo se equivoca de manera grave tiene una fuerte penalización, si no aplica una penalización menor o premia al modelo.

$$J(\beta) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right]$$

Donde:

- m: número de observaciones (541 pacientes)
- $y(i)$: etiqueta real (0 o 1)
- $\hat{y}(i)$: probabilidad predicha por el modelo

Método de Optimización - Descenso del Gradiente

El descenso del gradiente es el algoritmo que nos permite entrenar nuestro modelo de regresión logística, es decir, encontrar los mejores valores para los coeficientes β. Podemos entenderlo en cuatro pasos sencillos:

1. Empezamos con valores aleatorios para los coeficientes β
2. Calculamos la dirección de mejora (gradiente) que nos indica cómo ajustar los parámetros
3. Actualizamos los coeficientes moviéndonos en la dirección opuesta al gradiente
4. Repetimos el proceso hasta que el modelo deje de mejorar

¿Por qué funciona tan bien en regresión logística?

La clave está en que la función de costo que usamos (Log-Loss) tiene forma de tazón - es lo que llamamos una función convexa. Esto significa que no importa dónde empecemos, siempre vamos a llegar al mismo punto mínimo. No nos atascamos en mínimos locales que no son la mejor solución.

Además, tenemos control sobre qué tan grandes son nuestros pasos mediante la tasa de aprendizaje α. Si elegimos un valor muy pequeño, aprendemos lento pero seguro; si es muy grande, podemos pasarnos del óptimo. En nuestro caso, el algoritmo encontró el equilibrio perfecto para converger a una solución estable.

Pros y Contras Matemáticos

Este algoritmo es computacionalmente eficiente, el tiempo que tarda en entrenarse crece de manera predecible con el tamaño de nuestros datos, lo que la hace práctica incluso con datasets considerables.

Los coeficientes que obtenemos son fáciles de interpretar: cada uno nos dice exactamente cómo influye una variable en la probabilidad de tener PCOS. Por último, es numéricamente estable – no tenemos que preocuparnos demasiado por los valores iniciales que elijamos para los parámetros.

Limitaciones matemáticas:

Por otro lado, la regresión logística asume que la relación entre las variables y el resultado es lineal en escala logarítmica, lo que a veces es una simplificación de la realidad.

También es sensible a valores atípicos – si tenemos pacientes con características muy extremas, pueden afectar desproporcionadamente los coeficientes del modelo.

Otro problema potencial ocurre cuando las clases son fácilmente separables – en estos casos, los coeficientes pueden crecer demasiado y hacer que el modelo sea menos estable.

Finalmente, no captura relaciones complejas entre variables por sí solo. Si el efecto de una variable depende de otra (interacciones), necesitaríamos especificarlas manualmente.

¿Por qué la Regresión Logística es apropiada para diagnosticar PCOS?

En medicina, rara vez tenemos certezas absolutas – trabajamos con probabilidades. Que el modelo nos dé un valor entre 0 y 1 (como un 85% de probabilidad de PCOS) se parece mucho a cómo piensan los doctores en la vida real, evaluando riesgos en lugar de dar diagnósticos categóricos.

Cada coeficiente nos dice exactamente cuánto aporta cada síntoma al diagnóstico. Por ejemplo, que el "Follicle No. (R)" tenga un coeficiente de 1.093 nos confirma algo que los médicos ya saben: los folículos son un predictor muy importante. Esto genera confianza en el modelo.

Con la regularización que aplicamos (C=0.1), el modelo evita memorizar los datos de entrenamiento y se enfoca en aprender patrones generales. Esto significa que funcionará bien con pacientes nuevos, no solo con los que ya vimos.

Con 541 pacientes, tenemos suficiente información para entrenar un modelo confiable sin caer en sobreajuste. A diferencia de modelos más complejos que necesitan miles de muestras, la regresión logística nos da buenos resultados con los datos disponibles.

Aplicación Práctica: Implementación y Resultados

Configuración del modelo
Después de probar diferentes combinaciones, encontramos que estos ajustes funcionaron mejor para nuestro caso:

- C = 0.1 - usa una regularización moderada para evitar sobreajuste

- class_weight = 'balanced' - ayuda al modelo a aprender mejor de la clase minoritaria (pacientes con PCOS)
- solver = 'lbfgs' - el método de optimización más eficiente para nuestro dataset

Variables con mayor influencia en el diagnóstico

Al analizar los coeficientes del modelo, identificamos que estas 5 características son las que más aportan al diagnóstico de PCOS:

Follicle No. (R) : +1.0933

Follicle No. (L) : +0.6350

Hair growth (Y/N) : +0.5730

Skin darkening (Y/N) : +0.4847

Cycle (R/I) : +0.4538

Interpretación de los resultados

Tomemos como ejemplo el folículo del ovario derecho ($\beta = 1.0933$). Al calcular su odds ratio obtenemos 2.984, lo que significa que por cada folículo adicional, la probabilidad de tener PCOS se triplica aproximadamente. Esto confirma lo que sabemos médicamente: el número de folículos es un indicador clave del síndrome.

Visualización de los hallazgos

Creamos una gráfica de barras horizontales donde cada barra representa una variable y su longitud muestra qué tan importante es para el modelo. Esto nos permite ver de un vistazo qué síntomas tienen mayor peso en el diagnóstico (gráficas en el código realizado).

Análisis del desempeño del modelo

Usamos 0.5 como punto de corte para clasificar a los pacientes, solo cometimos 4 falsos negativos (pacientes con PCOS que el modelo no detectó) y 3 falsos positivos (pacientes sanos clasificados erróneamente). Estos números son bastante bajos, lo que muestra que nuestro modelo es confiable para apoyar el diagnóstico médico.

Discusión

Al juntar la teoría con lo que vimos en la práctica, notamos cosas muy interesantes. La propiedad matemática de que la función de costo tiene forma de tazón hizo que nuestro modelo fuera estable y confiable, pues los resultados fueron muy similares tanto con los datos de entrenamiento como con los nuevos datos de prueba.

Otra ventaja práctica fue que podemos entender fácilmente cómo toma decisiones el modelo. Por ejemplo, confirmó algo que los médicos ya sabían: que el número de folículos es el factor más importante para detectar PCOS. Esto le da más credibilidad a nuestros resultados.

También nos dimos cuenta de que el ajuste de regularización que aplicamos funcionó muy bien, ya que el modelo aprendió patrones generales en lugar de memorizar los datos. Esto se ve en que su capacidad para clasificar (AUC) fue casi igual de buena con pacientes nuevos (0.9589) que con los que ya conocía (0.9709).

Eso sí, hay que ser honestos sobre las limitaciones. Este modelo no detecta por sí solo relaciones complejas entre variables, como cuando el efecto de un síntoma depende de la presencia de otro. Además, asume que cada paciente es independiente, pero en la vida real podrían haber factores familiares o ambientales que relacionen a varios pacientes entre sí.

Conclusiones

La regresión logística se identifica como el mejor modelo para la clasificación de pacientes con Síndrome de Ovario Poliquístico (PCOS) debido a su superior desempeño global, con una exactitud del 93.58%, un F1-Score de 90.14% y un AUC-ROC de 0.9619, que reflejan un excelente equilibrio entre precisión y sensibilidad. Además, presenta un alto recall (88.89%) para la detección de casos positivos, lo cual es fundamental en un contexto clínico para minimizar falsos negativos y permitir intervenciones tempranas. Su interpretabilidad, mediante coeficientes que indican la influencia de cada variable, facilita la comprensión y aceptación por parte del personal médico. Finalmente, su robustez y estabilidad, evidenciadas por métricas consistentes en entrenamiento y prueba, junto con su eficiencia computacional, la convierten en la opción más adecuada para su implementación práctica en el diagnóstico asistido de PCOS.

Referencias

Fuente de Datos

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/