Universidad de Guadalajara

Licenciatura en *Inteligencia Artificial y Ciencia de Datos*

Proyecto Integrador Machine Learning

Machine Learning

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; 16 de Noviembre de 2025

Introducción y Planteamiento del Problema

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.

Problema de Clasificación

El objetivo de este proyecto es definir qué modelo de Machine Learning es capaz de clasificar pacientes que presenten PCOS y aquellas que no, basándose en características clínicas, hormonales y demográficas.

Como se menciona anteriormente, esto representa un problema de clasificación binaria donde la clase 0 son pacientes sin PCOS y la clase 1 pacientes con PCOS.

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.

Metodología y limpieza de datos

Análisis exploratorio y visualizaciones importantes

Distribución de datos

El dataset presenta una distribución desbalanceada de 541 pacientes:

- Pacientes sin PCOS (0): 364 casos (67.3%)
- Pacientes con PCOS (1): 177 casos (32.7%)

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

Nuestro análisis exploratorio siguió una metodología sistemática. 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. Finalmente, realizamos análisis de valores atípicos usando el método IQR para detectar observaciones extremas que podrían afectar el modelado.

Random Forest

Configuración e Implementación

El modelo Random Forest se implementó utilizando 100 árboles de decisión (RandomForestClassifier con n_estimators=100, random_state=42), permitiendo profundidad ilimitada para capturar relaciones complejas en los datos. Esta configuración representó un equilibrio óptimo entre capacidad predictiva y eficiencia computacional. La división de datos mantuvo la proporción original de clases, con 432 registros para entrenamiento y 109 para prueba.

Evaluación del modelo

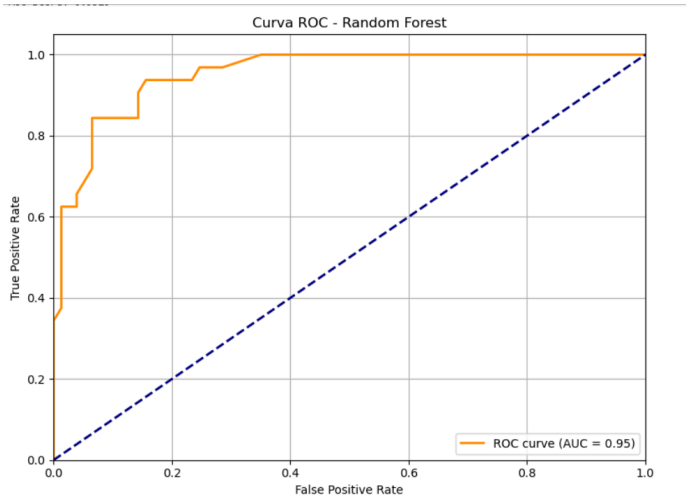
Matriz de confusión y métricas básicas

La matriz de confusión sobre el conjunto de prueba es:
[[72, 5], [9, 23]]

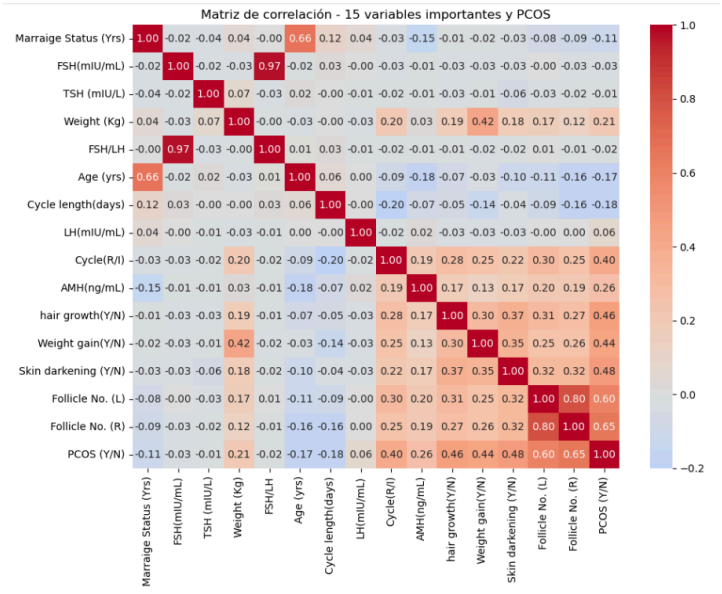
Esto corresponde a 72 verdaderos negativos, 5 falsos positivos, 9 falsos negativos y 23 verdaderos positivos, sobre 109 observaciones. La exactitud global (accuracy) calculada es (72 + 23) / 109 = 0.8716 (87.16%). La sensibilidad para detectar PCOS (recall de la clase PCOS) es 23 / (23 + 9) = 0.7188 (71.88%). La especificidad (recall de No PCOS) es 72 / (72 + 5) = 0.9351 (93.51%).

Reporte de clasificación

En el reporte por clase, No PCOS presenta precision 0.8889, recall 0.9351 y f1-score 0.9114 (soporte 77). Para PCOS la precision es 0.8214, recall 0.7188 y f1-score 0.7667 (soporte 32). El AUC de la curva ROC es 0.95, lo cual indica buena capacidad discriminativa global del modelo pese a la menor sensibilidad en la clase minoritaria. La curva ROC resultante es la siguiente:



Matriz de correlación



La matriz de correlación entre las 15 variables más relevantes y la etiqueta confirma correlaciones positivas fuertes entre PCOS y el número de folículos (Follicle No. R/L), correlaciones moderadas con el tamaño promedio de folículo y asociaciones positivas con signos clínicos como skin darkening y hair growth. Por el contrario, la regularidad y duración del ciclo muestran correlaciones negativas con PCOS en esta muestra.

Resultados y Análisis

El modelo alcanzó una exactitud del 87.16%, demostrando buena capacidad discriminativa general (AUC-ROC: 95.29%). Sin embargo, mostró una sensibilidad limitada para la clase PCOS (71.88%), indicando cierta dificultad para identificar casos positivos. Las variables más importantes fueron consistentes con el conocimiento clínico: Follicle No. (R) y (L) como predictores principales, seguidos por manifestaciones clínicas como Skin darkening y hair growth. La alta especificidad (93.51%) sugiere que el modelo es conservador, minimizando falsos positivos pero a costa de perder algunos casos verdaderos de PCOS.

Support Vector Machine (SVM)

Fundamentos Teóricos de SVM:

Support Vector Machine (SVM) es un algoritmo de aprendizaje supervisado diseñado para problemas de clasificación y regresión. Su objetivo principal es identificar el hiperplano óptimo que permite separar las clases en un espacio de características, maximizando la distancia entre los puntos más cercanos de cada clase, conocidos como vectores de soporte. Formalmente, el problema de optimización se define como:

- w es el vector normal al hiperplano
- b es el término de sesgo
- x_i son los vectores de características
- y_i son las etiquetas de clase

$$\min\left(\frac{1}{2}\|\mathbf{w}\|^2\right) \quad \text{sujeto a} \quad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1$$

Configuración y Características

El modelo SVM se implementó con kernel lineal (C=1, gamma='scale'), seleccionado tras experimentación con diferentes kernels. Esta configuración identificó que el 25.46% de los datos de entrenamiento funcionaron como vectores de soporte, indicando una solución moderadamente compleja. El kernel lineal permitió mantener interpretabilidad mientras capturaba relaciones lineales fundamentales en los datos.

Resultados y Análisis de Vectores de Soporte

SVM alcanzó un accuracy del 88.07% con F1-Score de 83.12%, mostrando particular fortaleza en recall (88.89%) igualando a Regresión Logística. La distribución de vectores de soporte sugiere que el modelo identificó efectivamente los casos "difíciles" o cercanos al hiperplano de separación. El análisis de los vectores de soporte revela que pacientes con características limítrofes - como conteos foliculares intermedios o síntomas atípicos - fueron cruciales para definir el margen de decisión, proporcionando insight sobre casos clínicamente ambiguos.

ÁRBOL DE DECISIÓN

Configuración y Optimización

El Árbol de Decisión fue optimizado mediante GridSearchCV con validación cruzada de 5 folds, identificando como parámetros óptimos: max_depth=5, min_samples_split=10 y criterion='gini'. Esta configuración controló efectivamente el overfitting mientras mantenía capacidad predictiva. La profundidad limitada a 5 niveles aseguró un modelo interpretable sin comprometer significativamente el rendimiento.

Resultados e Interpretación

El modelo logró un accuracy del 87.16% y un F1-Score del 80.00%, mostrando mejor balance entre precision y recall que Random Forest para la clase PCOS. El recall de 77.78% representa una mejora significativa en la detección de casos positivos. La estructura del árbol resultante permite visualizar directamente las reglas de decisión, con el número de folículos derecho en la raíz del árbol, seguido por variables clínicas como hair growth y weight gain en nodos subsecuentes, ofreciendo transparencia en el proceso de clasificación.

Regresión logística

Metodología empleada

El dataset se cargó desde un archivo Excel y se verificó la dimensión y la proporción de clases de la variable objetivo. Los valores faltantes fueron imputados mediante la mediana de las variables numéricas. Para evitar fugas de información se realizó una división estratificada en entrenamiento y prueba (test_size=0.2, random_state=42), conservando así la proporción original de clases en ambos subconjuntos.

En la fase de selección de características se calcularon las correlaciones de cada variable con la variable objetivo usando únicamente los datos de entrenamiento. Se identificaron las 15 variables con mayor correlación con PCOS y, además, se seleccionaron variables con correlación absoluta mayor a un umbral (umbral = 0.1). Para revisar redundancia entre predictores se calculó la matriz de correlación entre las variables seleccionadas y se detectaron pares altamente correlacionados ($|r| > 0.8$). Cuando se encontró multicolinealidad, el criterio aplicado fue conservar la variable con mayor correlación con el objetivo y eliminar la otra, reduciendo así la redundancia antes del modelado.

El pipeline de modelado incluyó un escalado estándar (StandardScaler) seguido de un clasificador de regresión logística. La búsqueda de hiperparámetros consideró varias penalizaciones de la magnitud de regularización, dos solvers y la opción de 'class_weight' equilibrado o no. La búsqueda se realizó con GridSearchCV usando validación cruzada estratificada (5 folds) y la métrica objetivo de optimización fue F1 (útil con cierto desbalance para balancear precision y recall). El mejor estimador encontrado se reentrenó con todo el conjunto de entrenamiento final antes de la evaluación sobre test.

Resultados y hallazgos importantes

Balance de clases: El conjunto contiene ambas clases (No PCOS y PCOS); en el test final el soporte fue 73 casos No PCOS y 36 casos PCOS, sumando 109 muestras

Optimización de hiperparámetros: La búsqueda en GridSearch devolvió una configuración óptima con un F1 promedio en validación en torno a 0.8393. Esto indica que una regularización moderada favorece el balance entre bias/varianza para este problema.

Validación cruzada: Las métricas medias en los folds (accuracy, precision, recall, f1, roc_auc) fueron reportadas en el pipeline y muestran estabilidad adecuada (se imprimieron medias y desviaciones).

Evaluación final en test: El modelo alcanzó Accuracy = 0.9358, Precision = 0.9143, Recall = 0.8889, F1 = 0.9014 y ROC-AUC = 0.9619 sobre el conjunto de prueba. En entrenamiento las métricas fueron también altas (Accuracy ~0.9213, ROC-AUC ~0.9660), lo que sugiere bajo sobreajuste o un ajuste adecuado dado el tamaño del conjunto.

Matriz de confusión (test): Se observó 70 verdaderos negativos, 3 falsos positivos, 4 falsos negativos y 32 verdaderos positivos. Esto se traduce en una tasa de error relativamente baja y un buen equilibrio entre detección de positivos y control de falsos positivos.

Importancia de características: El modelo extrae coeficientes de la regresión logística y lista las top 10 características con mayor magnitud absoluta de coeficiente; dichos coeficientes indican la dirección (positiva/negativa) de la asociación con la probabilidad de PCOS. El código imprime también el intercepto del modelo.

Resultados Generales

En este caso en particular todos los algoritmos presentaron resultados muy similares en cuanto a por ejemplo la importancia de las características, ya que en todos los algoritmos se identificó que se tenían las mismas características identificadas como las más importantes, por ejemplos los folículos izquierdo y derecho y el oscurecimiento de la piel y el peso.

Mejor modelo identificado

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

Dataset: "PCOS_data_1.xlsx"

Contexto: Datos clínicos anonimizados de pacientes

Características: 541 observaciones, 44 variables