TECNOCIENCIA CHIHUAHUA
Revista de ciencia y tecnología
ISSN-e: 2683-3360

UACH
UNIVERSIDAD AUTÓNOMA DE CHIHUAHUA

Artículo Científico

Predicción de ovario poliquístico aplicando técnicas de Machine Learning

Prediction of Polycystic Ovary Syndrome Applying Machine Learning Techniques

Carlos Eduardo Cañedo Figueroa[1*], Luisa Fernanda Blancarte Flores[1], Wendy Sofía Guerra Hernández[1], Daniela Licea Abúndez[1], Dafne Mariana Rivera Lerma[1] y Brianna Tena Holguín[1]

[1]Universidad Autónoma de Chihuahua, Facultad de Medicina y Ciencias Biomédicas. Circuito Universitario 31109, UACH Campus II, 31125 Chihuahua, Chih. México.

*Correspondencia: ccanedo@uach.mx (Carlos Eduardo Cañedo Figueroa)
DOI: https://doi.org/10.54167/tch.v17i2.1193
Recibido:: 30 de marzo de 2023; **Aceptado**: 19 de junio de 2023
Publicado por la Universidad Autónoma de Chihuahua, a través de la Dirección de Investigación y Posgrado

Resumen

El Síndrome de Ovario Poliquístico (SOP) es una de las endocrinopatías más comunes entre las mujeres que se encuentran en edad reproductiva. Hay estudios que exponen que esta patología afecta entre el 3-15 % de toda la población femenina. En el presente documento se describe el uso y la comparación de algunos algoritmos de machine learning con la finalidad de ofrecer una ventana de oportunidad en la clasificación de datos de forma eficiente. Por lo que se utilizaron tres algoritmos para realizar un diagnóstico del SOP contemplando 18 características extraídas de la base de datos "PCOS Dataset" alojada en la plataforma Kaggle.com. Se diseñaron una Red Neuronal Artificial (RNA) con un 97.5 % de F1, un algoritmo Bayesiano con un 97.6 % de F1 y un algoritmo de los K-Vecinos más Cercanos (KNN por sus siglas en inglés) con un 100 % de F1. El análisis realizado demostró que el algoritmo KNN clasifica los datos utilizados de forma óptima, lo que sugiere que puede ser utilizado para obtener diagnósticos en aplicaciones de laboratorio para obtener una evaluación complementaria.

***Palabras clave*: ovario poliquístico, red neuronal artificial, red Bayesiana, KNN, machine learning.**

Abstract

Polycystic Ovary Syndrome (PCOS) is one of the most common endocrinopathies among women of reproductive age. Studies show that this pathology affects between 3-15 % of the entire female

population. This paper describes the use and comparison of some machine learning algorithms with the aim of offering a window of opportunity in the classification of data in an efficient way. We used three machine learning algorithms to perform a diagnosis of the PCOS using 18 features extracted from the "PCOS Dataset" hosted on the Kaggle.com platform. An Artificial Neural Network (ANN) with 97.5 % F1, a Bayesian algorithm with 97.6 % F1 and a K-Nearest Neighbors (KNN) algorithm with 100 % F1 were designed. The analysis performed showed that the KNN algorithm classifies the data used optimally, suggesting that it can be used to obtain diagnostics in laboratory applications to obtain a complementary evaluation.

***Keywords*: polycystic ovary, artificial neural network, Bayesian network, KNN, machine learning.**

1. Introducción

El Síndrome de Ovario Poliquístico (SOP) es una de las endocrinopatías más comunes entre las mujeres que se encuentran en edad reproductiva (Guadamuz-Delgado, *et al*., 2022). Hay estudios que exponen que esta patología afecta entre el 3-15 % de toda la población femenina (Mubasher-Hassan, *et al*., 2020). La principal causa de este trastorno es una anormalidad en los ovarios, pero algunos agentes adicionales tales como el sobrepeso y factores ambientales pueden influir en el desarrollo de los síntomas individuales del SOP (Aguayo-González, 2023). Actualmente se han estado utilizando los Criterios de Rotterdam (2003) para su diagnóstico. Este trastorno es diagnosticado si se cumplen dos de las tres condiciones que presentan estos criterios: 1) Hiperandrogenismo clínico o bioquímico, 2) Anormalidades en la ovulación (Oligoovulación crónica), y 3) Poliquistosis ovárica por ecografía y un volumen ovárico mayor a 10 ml (Carvajal, et al., 2010).

Esta anomalía es un trastorno endocrino que se diagnostica después de descartar otras patologías con síntomas similares, como pueden ser la hiperplasia suprarrenal congénita no clásica, tumores productores de andrógenos, el síndrome de Cushing y otras formas de hiperandrogenismo. Esto hace que su diagnóstico sea complejo, ya que existe una alta heterogeneidad de su expresión clínica, Esto toma importancia debido a comorbilidades metabólicas y trastornos reproductivos (Mubasher Hassan, 2020).

La evidencia sugiere que el hiperandrogenismo es el factor más determinante en la fisiopatología del SOP, lo que se puede determinar si se observan síntomas como hipertensión, acné, menstruación irregular y producción inmoderada de andrógenos. Cabe destacar que el SOP es una de las principales causas de infertilidad femenina, ya que impide la correcta evolución de los folículos. (Winnykamien *et al*., 2016).

Actualmente se ha reportado el uso automatizado de imágenes de ultrasonido para la detección de SOP y hay algunos trabajos de Machine Learning (Alam-Suha *et al*., 2022). En este trabajo fueron utilizados tres métodos de Machine Learning: Clasificador Bayesiano Ingenuo (NB), K-Nearest Neighbors (KNN) y Red Neuronal Artificial (ANN). Para poder hacer uso de estos fue necesario elegir las principales características responsables del SOP y gracias a ellas lograr crear un modelo predictivo para la identificación del SOP con la finalidad de generar un algoritmo que sirva como apoyo a los especialistas de salud y con ello evitar el error humano.

2. Materiales y Métodos

2.1 Materiales

Para el desarrollo de este proyecto fue necesario tener acceso a la plataforma de datos que se encuentra en la página de Kaggle.com (Moheddine, 2022). Para el desarrollo de los métodos se utilizó una versión de Matlab con licencia de estudiante en un equipo portátil de la compañía HP con procesador Intel Core i7 y 8 gigabytes de memoria RAM bajo el sistema operativo de Windows 11 para 64 bits.

2.2 métodos

Procesar y seleccionar correctamente los atributos más característicos del SOP fue indispensable para obtener información fundamentada con el objetivo de ofrecer un respaldo estadístico ante un diagnóstico realizado por un médico.

Parte de la metodología incluyó la modificación de una base de datos existente y la elaboración de distintos algoritmos de Machine Learning: Clasificador Bayesiano Ingenuo (NB), Red Neuronal Artificial (ANN), y el Algoritmo del Vecino más Cercano (KNN). Después de realizar los diferentes algoritmos, con el objetivo de comprobar que los métodos aplicados fueron adecuados, se realizó un análisis de tipo F1 Score donde se evalúa la veracidad de todos los algoritmos.

2.3 Base de datos

La base de datos (DB) utilizada fue la PCOS_Dataset (Moheddine, 2022) que contiene el historial clínico de 541 pacientes diagnosticados con o sin SOP. Los datos son de libre uso dentro de la plataforma Kaggle.com y tienen oculta la información sensible de los participantes. Por lo que solo se puede tener acceso a la información de sus variables respecto al SOP y las clases a las que corresponden. Fue necesario hacer una reorganización y filtración de los datos ya que se contaba con varios diagnósticos y características como el grupo sanguíneo, si presentaba un embarazo, si había practicado un aborto anteriormente y cuántos, en caso de estar en un matrimonio, cuántos años tenía, si presentó algún aumento de peso y medidas morfológicas. Así como características de sus hábitos, tales como la actividad física o el consumo de comida chatarra. Por lo que, DB se focalizó en 18 de 44 características, utilizando sólo aquellas que poseen información numérica distinta de valores 1 y 0 ya que esto podría afectar el desarrollo de algoritmos. En la Tabla 1 se enlistan las 18 que fueron utilizadas.

Si bien los datos obtenidos de DB contiene 541 registros, se utilizaron 237 registros (68 de la clase PCOS + y 169 de la clase PCOS -), los cuales contienen las características indicadas en la Tabla 1 de forma completa, el resto de los registros, indicaban espacios vacíos, por lo que fueron descartados. A su vez, se dividió en cuatro partes de forma aleatoria: 65 % de la clase positiva (BDP), 65 % de la clase negativa (BDN), el 35 % restante de la clase positiva (BDTP) y el 35 % restante de la clase negativa (BDTN) esto para obtener datos con los cuales realizar pruebas después del entrenamiento de los algoritmos generados.

Tabla 1. Características utilizadas de la base de datos.
Table 1. Used database characteristics.

No.	Característica
1	Índice de Masa Corporal (BMI)
2	Hemoglobina
3	Gonadotrofina Coriónica Humana I (I β-HCG)
4	Gonadotropina Coriónica Humana II (II β-HCG)
5	Hormona Estimulante del Folículo (FSH)
6	Hormona Luteinizante (LH)
7	Relación FSH/LH
8	Hormona Estimulante de la Tiroides (TSH)
9	Hormona antimülleriana (AMH)
10	Prolactina (PRL)
11	Vitamina D3
12	Progesterona (PRG)
13	Glucosa en Sangre (RBS)
14	Presión Sistólica
15	Presión Diastólica
16	Cantidad de folículos en el ovario izquierdo
17	Cantidad de folículos en el ovario derecho
18	Tamaño del Endometrio

2.4 Clasificador bayesiano ingenuo

Para el desarrollo del primer algoritmo, se optó por un clasificador Bayesiano Ingenuo (NB), mediante el cual se le asigna una etiqueta a un elemento según sus características, rasgos o propiedades, mediante el análisis de los datos basado en la ecuación Ec. (1) del teorema de Bayes.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad \text{Ec. (1)}$$

Donde "A" y "B" son eventos aleatorios. Siendo la probabilidad del evento "A" en el evento "B" en términos de la distribución de la probabilidad condicional del evento "B" en el evento "A"

multiplicado por la distribución de probabilidad marginal de "A" (Palomar, *et al*., 2017). El algoritmo se puede describir tal cual como se muestra en la Fig. 1.

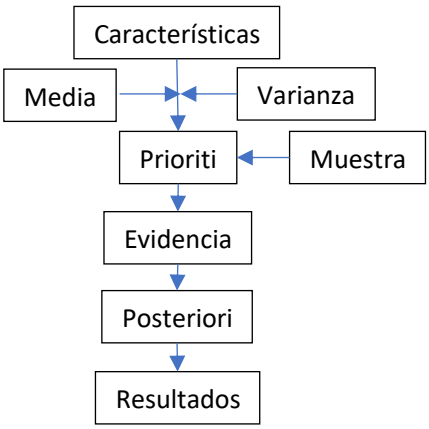

Figura 1. Diagrama de flujo del algoritmo Bayesiano.
Figure 1. Diagram of Bayesian algorithm.

Una vez realizados los cálculos se tomó en cuenta la condición que se puede observar en la Ec. (2) para la clasificación del dato muestra, en donde *Z(1)* hace referencia a la probabilidad de que el vector característico sea positivo y *Z(2)* a la probabilidad obtenida del vector como negativo.

$$Clase = \begin{cases} positivo & si \quad Z(1) > Z(2) \\ negativo & si \quad Z(2) \geq Z(1) \end{cases} \qquad \text{Ec. (2)}$$

La variable *Clase* podría quedar dentro de una de las dos clases tomando en cuenta los intervalos calculados: si toma el valor de -*positivo*-, entonces, los datos ingresados provienen de un paciente con SOP, en cambio si el resultado se encontró fuera del intervalo, los datos pertenecen a un paciente sin SOP.

2.5 Red neuronal (ANN)

Una red neuronal artificial es el conjunto de secuencias matemáticas que pueden recibir datos a modo de características, cuentan con una función de activación, una polarización y sus respectivos puntos sinápticos, que al momento de ser entrenada por medio de diversos algoritmos puede clasificar un dato muestra (García-Chavez *et al*., 2021).

La estructura de la red neuronal (ver Fig. 2) fue base para el desarrollo del siguiente algoritmo de clasificación, se entrenó con los siguientes híper parámetros de forma experimental:

- factor de aprendizaje de 0.01,
- 3,000 épocas
- 5 neuronas
- Una capa oculta,
- Verificación de mínimos locales de 2,000
- Error máximo 1e-25.

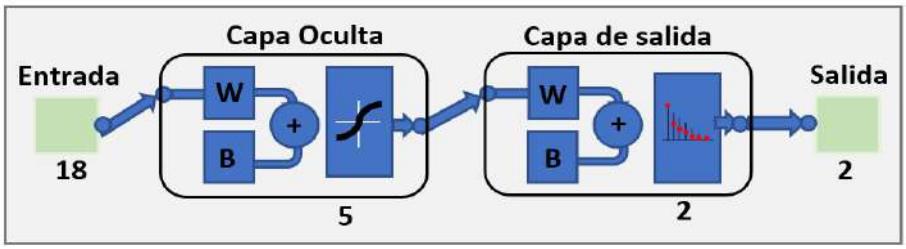

Figura 2. Estructura de la Red Neuronal Artificial.
Figure 2. ANN estructure.

Para el entrenamiento de la red neuronal se utilizó el algoritmo de gradiente descendiente basado en retro propagación utilizando *BDP* y *BDN* con las 18 características de la Tabla 1 como conjuntos de entrenamiento.

Para poder hacer uso de la red neuronal entrenada, se utilizó la Ec. (3), considerando que $Muestra_i$ es el vector característico por evaluar. La matriz de confusión obtenida después del entrenamiento quedo como se muestra en el Fig. 3.

$$y = ANN(Muestra_i) \qquad \text{Ec. (3)}$$

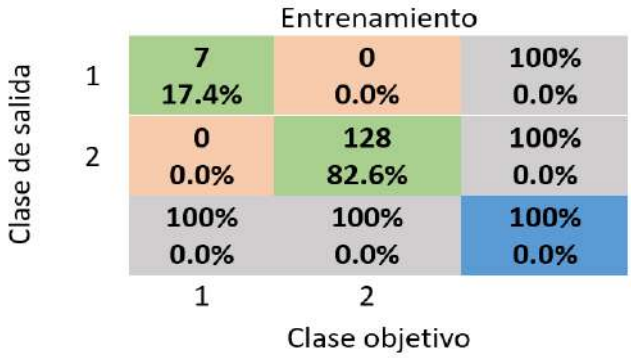

Figura 3. Matriz de confusión de la red neuronal.
Figure 3. Confusion matrix of ANN.

K vecinos cercanos (KNN)

El algoritmo del vecino más cercano es una de las formas más sencillas de Machine Learning para poder clasificar una base de datos usando un modelo de entrenamiento tomando las distancias euclidianas entre los datos vecinos que se encuentren más cercanos; dándose a sí mismo opciones para su propia modificación con el objetivo de reducir las limitaciones y obstáculos, así como mejorar su precisión y aplicabilidad (Uddin *et al.*, 2022).

Se realizó una comparación de las distancias obtenidas para efectuar la clasificación adecuada y mediante la función "fitcknn" ejecutada en Matlab®, se programó el KNN con las 18 variables de la Tabla 1.

3. Resultados y discusión

Las métricas de los resultados obtenidos por cada algoritmo se obtuvieron mediante las ecs. (4)-(7) con lo que se logró obtener la Precisión, Recall, Exactitud y F1 score. En donde *TP (True Positive)* son los verdaderos positivos, *TN (True Negative)* verdaderos negativos, *FN (False Negative)* falsos negativos y *FP (False Positive)* falsos positivos, obtenidos al pasar los datos de *BDTP* y *BDTN* por los algoritmos desarrollados.

$$\text{Precisión} = \frac{TP}{FP + TP} \qquad \text{Ec. (4)}$$

$$\text{Recall} = \frac{TP}{TP + FN} \qquad \text{Ec. (5)}$$

$$\text{Exactitud} = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{Ec. (6)}$$

$$\text{F1} = 2\left(\frac{\text{Precisión} \bullet \text{Recall}}{\text{Precisión} + \text{Recall}}\right) \qquad \text{Ec. (7)}$$

Al analizar las métricas de los resultados de cada algoritmo desarrollado en conjunto con los algoritmos de otros autores se obtuvieron los datos expuestos en la Tabla 2, en los que la precisión máxima fue obtenida mediante el algoritmo de aprendizaje supervisado KNN con un valor del 100 %. Con un F1 score de 97.5 % para el algoritmo bayesiano, 97.6 % para ANN y 100 % para KNN.

Tabla 2. Comparación de resultados obtenidos.
Table 2. Comparison of results obtained.

	Precisión	**Recall**	**Exactitud**	**F1 Score**
Bayesiano	95.1 %	100 %	97.6 %	97.5 %
ANN	97.6 %	97.6 %	98.8 %	97.6 %
KNN	100 %	100 %	100 %	100 %
Bayes (Prapty y Shitu, 2020)	80 %	80 %	93 %	80 %
Random Forest (Prapty y Shitu, 2020)	84 %	84 %	93.5 %	84 %
KNN (Denny *et al*., 2019)	83.33 %	-	86.58 %	40.98 %
KNN (Madhumitha *et al.,* 2021)	97 %	95 %	97 %	95.98 %
Bayes (Mubasher-Hassan y Mirza, 2020)	94 %	76 %	81 %	84 %

Los datos expuestos en la Tabla 2 sugieren que el uso de algoritmos de agrupamiento basadas en las 18 características que se utilizaron en el presente estudio son óptimos para la detección y clasificación de SOP. Superior 2.4 % en comparación con la red neuronal desarrollada y a 2.5 % del algoritmo Bayesiano desarrollado.

Se puede denotar una diferencia en la puntuación del F1 score con los autores (Prapty *et al*., 2020) quienes utilizaron características booleanas como el ciclo menstrual regular, grosor del cabello, pérdida de cabello y la denotación de cambios físicos en el cuerpo. De igual manera (Denny *et al*., 2019) trabajo con datos que pueden ser considerados como booleanos dentro de su conjunto de datos.

(Madhumitha *et al*., 2021) y (Mubasher Hassan *et al*., 2020) quienes utilizaron datos obtenidos de imágenes, lograron generar algoritmos competitivos, sin embargo, se puede notar una diferencia en el F1 score de sus métricas con las generadas en el presente documento.

Tras una comparación de los resultados de los distintos trabajos con el presente, es posible identificar una gran disparidad de valores con una diferencia de 3 % hasta 20 % en la precisión, de 5 % a 20 % en el recall, de 3 % a 19 % en la exactitud y de 16 % hasta 59 % en el F1 score. Los resultados obtenidos sugieren que el desarrollo de algoritmos de machine learning tienen un mejor desempeño cuando se involucran datos no booleanos dentro de entrenamiento incrementado las métricas que se puedan obtener.

4. Conclusiones

Los resultados obtenidos, muestran que en términos generales es recomendable utilizar características descriptivas no booleanas para el desarrollo de algoritmos de machine learning. De igual manera se deberán validar los algoritmos generados utilizando información de algunas otras anomalías similares para obtener una evaluación complementaria de lo desarrollado en este documento.

Es necesario destacar que la presencia del Síndrome de Ovario Poliquístico en la salud femenina ha alcanzado un importante interés dentro de los factores de riesgo para las mujeres en edad reproductiva, puesto que es una anomalía que afecta directamente en los problemas de infertilidad y de salud materna.

Se puede concluir que el uso de inteligencia artificial para el diagnóstico de este tipo de enfermedades puede ser muy útil. Por lo tanto, se puede generar alguna aplicación que resulte práctico en laboratorios para el pre diagnóstico de la existencia del SOP, así como también el desarrollo de herramientas de asistencia al médico.

Agradecimientos

Agradecemos a la Universidad Autónoma de Chihuahua y a los creadores del repositorio de la base de datos Aya Moheddine y su equipo por la disposición de colocar su aportación en la plataforma de libre acceso Kaggle.com en pro del desarrollo científico y el bienestar de la humanidad.

Conflicto de intereses

Los autores de este artículo declaran no tener ningún conflicto de interés.

5. Referencias

Aguayo-González, P. 2016. Sobrepeso y obesidad, factores de riesgo para desarrollar síndrome de ovario poliquístico. https://www.gob.mx/salud/prensa/sobrepeso-y-obesidad-factores-de-riesgo-para-desarrollar-sindrome-de-ovario-poliquistico

Suha, S.A. & Islam, M.N. 2022. An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image. Nature Scientific Reports 12: 17123. https://doi.org/10.1038/s41598-022-21724-0 *

Cañedo-Figueroa, C. E. & García-Chávez, H. 2021. Diseño de algoritmo compuesto por Machine Learning y un modelo probabilístico para la detección de diabetes. Memorias Del Congreso Nacional de Ingeniería Biomédica 8(1), 57–60. https://memoriascnib.mx/index.php/memorias/article/view/828

Carvajal, R., Herrera, G. & Porcile, J., 2010. Espectro Fenotípico Del Síndrome De Ovario Poliquístico. Rev Chil Obstet Ginecol, 75(2): 124 – 132. http://dx.doi.org/10.4067/S0717-75262010000200009

Denny, A., Raj, A., Ashok, A., Maneesh-Ram, C. & George, R. 2019. i-HOPE: Detection and Prediction System for Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques. En: TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 673-678. https://doi.org/10.1109/TENCON.2019.8929674

Guadamuz-Delgado, J., Miranda-Saavedra, M. & Mora-Miranda, N. 2022. Actualización sobre el síndrome de ovario poliquístico. Revista Médica Sinergia 7(5): e801. https://doi.org/10.31434/rms.v7i5.801

Madhumitha, J., Kalaiyarasi, M. & Ram, S. S. 2021. Automated Polycystic Ovarian Syndrome Identification with Follicle Recognition. En 2021 3rd International Conference on Signal Processing and Communication (ICPSC), 98-102. https://doi.org/10.1109/ICSPC51351.2021.9451720

Moheddine, A., 2022. PCOS_Dataset Kaggle. https://www.kaggle.com/datasets/ayamoheddine/pcos-dataset?select=PCOS_data.csv

Mubasher-Hassan, M. & Mirza, T. 2020. Comparative Analysis of Machine Learning Algorithms in Diagnosis of Polycystic Ovarian Syndrome. Int J Comput Appl, 175(17), 42-53. https://www.ijcaonline.org/archives/volume175/number17/31548-2020920688

Palomar, L. & Guerrero, J. 2017. El teorema de bayes y el diagnóstico clínico. Memorias del Congreso Internacional Sobre la Enseñanza y Aplicación de las Matemáticas, Universidad Nacional Autónoma de México, Facultad de Estudios Superiores Cuautitlán. https://bit.ly/3JBftTU

Prapty, A. S. & Shitu, T. T. 2020. An Efficient Decision Tree Establishment and Performance Analysis with Different Machine Learning Approaches on Polycystic Ovary Syndrome. En: 23rd International Conference on Computer and Information Technology (ICCIT), DHAKA, Bangladesh, 2020, pp. 1-5. https://doi.org/10.1109/ICCIT51783.2020.9392666

Uddin, S., Haque, I., Lu, H., Moni, M. A. & Gide, E. 2022. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Sci Rep, 12(1): 6256. https://doi.org/10.1038/s41598-022-10358-x

Winnykamien, I., Dalibón, A. & Knoblovits, P. 2016. Síndrome de ovario poliquístico. Rev. Hosp. Ital. B. Aires 37(1): 10-20. https://pesquisa.bvsalud.org/portal/resource/pt/biblio-966680