UC | Facultad de Ciencias

INTELLCYST: HERRAMIENTA DE DIAGNÓSTICO BASADA EN IA PARA EL SÍNDROME DE OVARIO POLIQUÍSTICO
(INTELLCYST: AI-BASED DIAGNOSIS TOOL FOR POLYCYSTIC OVARY SINDROME)

Trabajo de Fin de Máster
para acceder al

MÁSTER EN CIENCIA DE DATOS

Autora: Paula Monje Ibáñez

Director\es: Diego Tuccillo
Alejandro González

Junio - 2024

Resumen

En este trabajo se ha desarrollado la aplicación Intellcyst, una herramienta de Inteligencia Artificial para el diagnóstico y tratamiento del Síndrome de Ovario Poliquístico (SOP). El SOP es la principal enfermedad reproductiva que sufren las mujeres, y su sintomatología es muy diversa, lo que en muchos casos dificulta el diagnóstico. Debido a esto se ha desarrollado un modelo de diagnóstico, utilizando distintos algoritmos de Aprendizaje Automático, y el conocimiento experto sobre el SOP aplicando un sistema híbrido. Así mismo, se ha diseñado una aplicación que permite al usuario utilizar este modelo de forma sencilla, para obtener una probabilidad de diagnóstico y un tratamiento personalizado y específico para cada paciente.

Palabras clave:
Síndrome de Ovario Poliquístico, diagnóstico, Inteligencia Artificial, Aprendizaje Automático, tratamiento

Abstract

This work presents the development of Intellcyst, an Artificial Intelligence tool for the diagnosis and treatment of the Polycystic Ovary Syndrome (PCOS). PCOS is the most common disease in women of preproductive age, and its symptomatology is very diverse, what can lead to difficulties in its diagnosis. Because of this we have decided to develop a diagnosis model, using different Machine Learning algorithms and the expert knowledge of PCOS in a hybrid system. In addition, we have designed an app that allows the user to access easily to this diagnosis model, in order to obtain a diagnosis probability, and a specific and personalized curse of treatment.

Keywords:
Polycystic Ovary Syndrome, diagnosis, Artificial Intelligence, Machine Learning, treatment

Índice general

Capítulo 1

Introducción

1.1. Contexto

El Síndrome de Ovario Poliquístico (SOP) es la enfermedad endocrina más común en mujeres en edad reproductiva, aunque se puede presentar también en la adolescencia y en la menopausia [1]. El SOP afecta a entre el 6 % y el 20 % de la población [2]. Se trata de una enfermedad compleja, ya que su etiología (causa) es desconocida, y la presentación clínica es muy heterogénea, presentando síntomas metabólicos, reproductivos y psicológicos. Así mismo, el SOP es la primera causa de infertilidad de las mujeres. Debido a esto, el diagnóstico es complicado, y muchas veces se produce de forma tardía. Así mismo, las formas de diagnóstico y tratamiento son inconsistentes entre países, y en consecuencia muchas mujeres sufren un diagnóstico tardío y un tratamiento insuficiente.

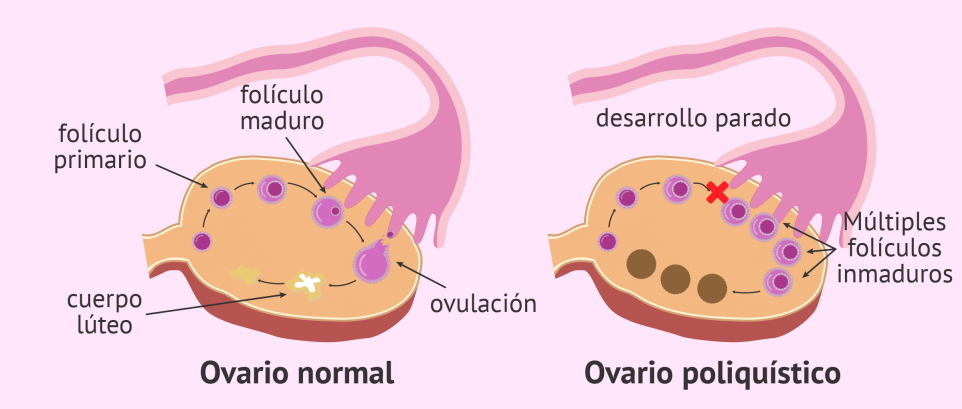

Figura 1.1: Ovario normal vs. ovario poliquístico [3].

Como la sintomatología del SOP es tan diversa, durante muchos años ha sido un reto crear un consenso de diagnóstico correcto. Es debido a esto que muchas mujeres sufren un retraso significativo en su diagnóstico, sobre todo en aquellas comunidades donde el acceso al sistema sanitario es más difícil.

1.2. Motivación

En España, entre 1 y 5 millones de mujeres tienen Síndrome de Ovario Poliquístico. La mayoría de estas mujeres sufren síntomas que afectan a su vida diaria, y muchas tendrán problemas de fertilidad si intentan quedarse embarazadas. Muchas de estas mujeres tardarán meses o incluso años en recibir un diangóstico certero por parte del personal ginecológico. Muchas de estas mujeres, cuando reciban el diagnóstico, se someterán a un proceso de pérdida de peso, a

un tratamiento farmacológico para limitar sus síntomas, o un tratamiento de fertilidad. Tendrán que gestionar una enfermedad que no tiene cura durante la mayor parte de su vida. Y en otras partes del mundo donde el acceso a consultas médicas y tratamientos no está disponible a todo el mundo, muchas mujeres se pasarán toda su vida sufriendo estos síntomas, sin saber siquiera que tienen SOP. Es debido a todo esto que hemos decidido proponer una posible solución a este problema.

1.3. Objetivos

El principal objetivo de este trabajo es desarrollar un modelo de diagnóstico en tres etapas que sea capaz de diagnosticar correctamente el Síndrome de Ovario Poliquístico. La idea es obtener un modelo que sea capaz de diagnosticar el SOP sin necesidad de hacer pruebas médicas, o de que la paciente espere a una consulta médica. De esta forma el primer objetivo es que la paciente obtenga un diagnóstico lo bastante certero utilizando esta aplicación, o que el personal sanitario sea capaz de diagnosticar a la paciente en la primera consulta a la que acuda.

En el caso de que este diagnóstico no sea posible sin realizar pruebas médicas, el objetivo es que la aplicación sea capaz de dar un diagnóstico certero utilizando los resultados de estas pruebas médicas.

Así mismo, parte del objetivo de este trabajo es poder proporcionar un curso de tratamiento personalizado a las pacientes que sean diagnosticadas con SOP. Para esto se utilizará el conocimiento médico y las últimas publicaciones sobre el SOP para tratar de proporcionar el tratamiento adecuado.

1.4. Intellcyst

Ante este problema se ha propuesto como posible solución el desarrollo de la plataforma Intellcyst, diseñada para facilitar el diagnóstico del Síndrome de Ovario Poliquístico, así como para ofrecer un tratamiento y acompañamiento personalizado para cada paciente. Intellcyst utilizará modelos de Aprendizaje Automático para tratar de diagnosticar a las pacientes en tres fases diferentes del diagnóstico:

1. En una primera visita, donde solo se conocen los síntomas y la información básica de la paciente, sin hacer ninguna prueba médica. De esta forma un personal sanitario, puede proporcionar un diagnóstico certero en la primera consulta a la que atienda la paciente.

2. En los casos en los que la probabilidad del diagnóstico no supere un umbral determinado, se recomendará realizar una ecografía ovárica, y utilizando esta información se desarrollará un nuevo modelo, que produzca un diagnóstico más certero.

3. En los casos en los que la probabilidad de diagnóstico anterior tampoco supere determinado umbral, se recomendará realizar una analítica sanguínea, y utilizando esta información se desarrollará un nuevo modelo, que produzca un diagnóstico certero.

Intellcyst también proporcionará un curso de tratamiento personalizado para cada paciente. Utilizando el conocimiento médico, la sintomatología y las características específicas de cada paciente, se desarrollará un curso de tratamiento que abarque cuatro entornos principales: estilo de vida, tratamiento farmacológico, fertilidad y tratamiento psicológico.

En la Figura 1.2 se puede observar un esquema del funcionamiento de la aplicación que se pretende desarrollar.

Esta aplicación se ha desarrollado en el marco de mis prácticas de Máster en la empresa Deduce Data Solutions, y con la colaboración del doctor Alejandro González, en colaboración con la clínica Arpa Médica y el Grupo Fertilidad Oyala.

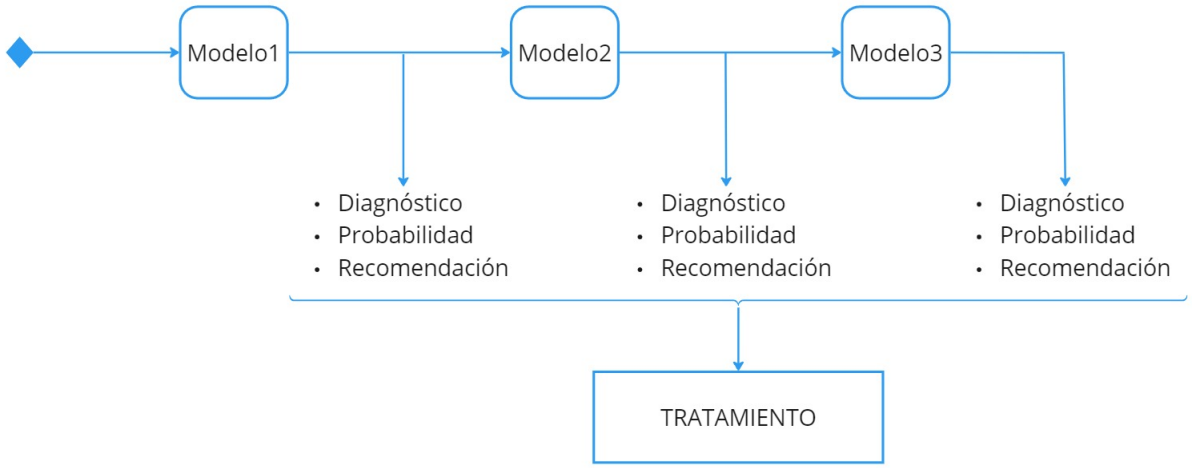

Figura 1.2: Esquema del funcionamiento de la aplicación Intellcyst.

1.5. Estructura

Para cumplir con estos objetivos, en primer lugar se realizará un estudio sobre el Síndrome de Ovario Poliquístico, para tratar de entender las características más importantes de esta enfermedad, así como revisar la literatura existente. Después se estudiarán las herramientas de Inteligencia Artificial que pueden resultar útiles en este trabajo, así como el estado del arte y otros ejemplos de aplicaciones de Inteligencia Artificial en el campo de la medicina. A continuación se analizarán los *datasets* utilizados para desarrollar esta herramienta, así como sus similitudes y diferencias. Seguidamente se explicará cómo se produjo el desarrollo del modelo, y cómo se aplicó el conocimiento médico a los algoritmos de Aprendizaje Automático utilizados. Seguidamente, se explicará el desarrollo del Producto Mínimo Viable (MVP), es decir el diseño de la aplicación Intellcyst y la integración de este modelo de diagnóstico en dicha aplicación. Finalmente, se terminará con las conclusiones del trabajo.

Capítulo 2

Síndrome de Ovario Poliquístico

2.1. Revisión de la literatura

Según la Guía Internacional Basada en Evidencia para la Evaluación y Manejo del Síndrome de Ovario Poliquístico [1], actualizada en 2023 y que sigue la recomendación anterior de 2018, los criterios para diagnosticar el SOP deberían ser los establecidos en el criterio de Rotterdam de 2003 [4]. Según este criterio, se requieren dos de los siguientes síntomas para diagnosticar el Síndrome de Ovario Poliquístico:

- Hiperandrogenismo clínico o bioquímico: exceso de hormonas masculinas o sus manifestaciones clínicas, como pueden ser el hirsutismo, pérdida de cabello, o acné.
- Oligo-anovulación o disfunción ovulatoria, que se presenta generalmente como irregularidad en el ciclo menstrual.
- Morfología poliquística en el ovario: un número excesivo de folículos preantrales (inmaduros).

Actualmente, en la guía de 2023 se incluye un nuevo síntoma que puede sustituir a la ecografía: el exceso de la hormona anti-Mülleriana (AMH). Es importante destacar que en los casos que se presenten ciclos menstruales irregulares e hiperandrogenismo, se puede diagnosticar directamente el Síndrome de Ovario Poliquístico, sin necesidad de hacer ecografías o analizar la AMH. Es más, en caso de pacientes adolescentes, se requiere solamente el hiperandrogenismo y la disfunción ovulatoria como síntomas, ya que las ecografías y el análisis de AMH no están recomendados en este caso. Una vez sea diagnosticado el SOP, se deben analizar sus posibles implicaciones en problemas reproductivos, metabólicos, cardiovasculares, dermatológicos, de sueño y psicológicos.

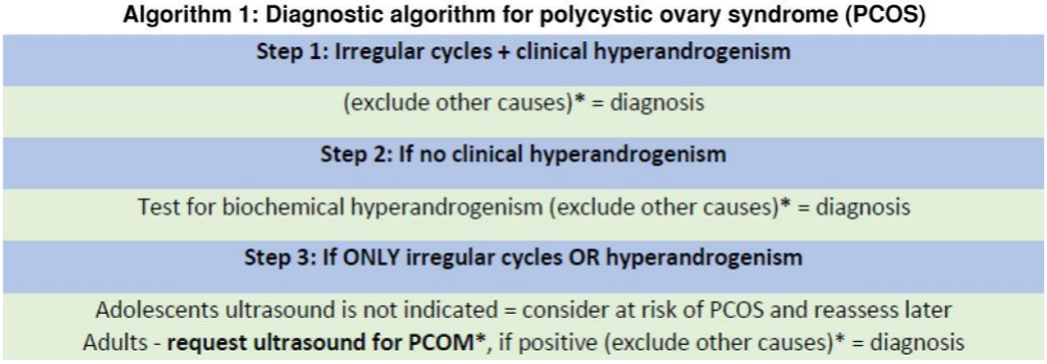

Figura 2.1: Algoritmo de diagnóstico del Síndrome de Ovario Poliquístico [4].

Hay muchos estudios que analizan la etiología y patología del Síndrome de Ovario Poliquístico. La mayoría señalan la causa en una mezcla de factores externos e internos que incluyen la resistencia a la insulina, el hiperandrogenismo, factores ambientales, genéticos y epigenéticos [5].

Así mismo, se ha buscado un origen genético en la enfermedad, aunque a día de hoy no se ha encontrado un único gen que se pueda relacionar con el SOP [6]. Se han encontrado genes, relaciones entre genes y relaciones de genes con el ambiente que pueden potenciar la aparición de la enfermedad, pero no se pueden relacionar con una causa directa. El SOP es fuertemente heredable, entre un 60 % y 70 % de las hijas de mujeres con SOP manifiestan la enfermedad en la adolescencia o la edad adulta [7]. Es debido a esto que se está estudiando el posible origen epigenético de la enfermedad.

Tampoco se puede afirmar que haya una línea de tratamiento clara y generalizada para el SOP. Depende en mayor medida de las características de la paciente, y de si su prioridad es buscar un tratamiento de fertilidad, la regulación menstrual o la pérdida de peso. El tratamiento debe ser personalizado para alcanzar un resultado óptimo [5].

Una primera línea de tratamiento suele ser la pérdida de peso, a través del ejercicio regular y una dieta saludable. En este aspecto se recomienda una dieta personalizada para cada paciente, que atienda a sus síntomas y características específicas. En cuanto al ejercicio, las recomendaciones también deben atender a la situación personal de la paciente, aunque la recomendación general es de 150 minutos a la semana de ejercicio moderado o 75 minutos de ejercicio vigoroso.

El tratamiento farmacológico mayormente recomendado a las mujeres con SOP es los anticonceptivos combinados orales (ACOs). Este tratamiento se suele recomendar a las mujeres que no se encuentran en un proceso de búsqueda del embarazo, y que buscan mejorar los síntomas del hiperandrogenismo clínico o la irregularidad del ciclo menstrual. Otro fármaco muy recomendado es la metformina, en aquellos casos de mujeres que presentan irregularidad en el ciclo ovulatorio debido a la resistencia a la insulina. En el caso de las mujeres con problemas de fertilidad, se puede realizar un tratamiento de inducción de la ovulación utilizando citrato de clomifeno. También se pueden recomendar tratamientos farmacológicos para la pérdida de peso, los problemas dermatológicos, el hirsutismo, etc.

2.2. Características del Síndrome de Ovario Poliquístico

El Síndrome de Ovario Poliquístico puede clasificarse en varios fenotipos, es decir, en varias manifestaciones de la enfermedad [8]. Estos fenotipos se dividen en función de los síntomas que presenta la paciente: el fenotipo A presenta los tres síntomas que se utilizan para el diagnóstico del SOP: presencia de ovario poliquístico, oligo-anovulación e hiperandrogenismo. El fenotipo B presenta oligo-anovulación e hiperandrogenismo, el fenotipo C presenta ovario poliquístico e hiperandrogenismo y el fenotipo D presenta ovario poliquístico y oligo-anovulación

Fenotipo	**PCO**	**OA**	**HA**
A	Sí	Sí	Sí
B	No	Sí	Sí
C	Sí	No	Sí
D	Sí	Sí	No

Tabla 2.1: Posibles fenotipos del Síndrome de Ovario Poliquístico. Los síntomas son, de izquierda a derecha: ovario poliquístico (PCO), oligo-anovulación (OA) e hiperandrogenismo (HA).

Aunque los síntomas del Síndrome de Ovario Poliquístico son más evidentes durante la etapa reproductiva, también se pueden presentar durante la prepubertad y la postmenopausia. Los síntomas en niñas pueden incluír una pubarquia prematura, síntomas tempranos de androgenización como acné e hirsutismo, e irregularidad menstrual. En las mujeres postmenopaúsicas

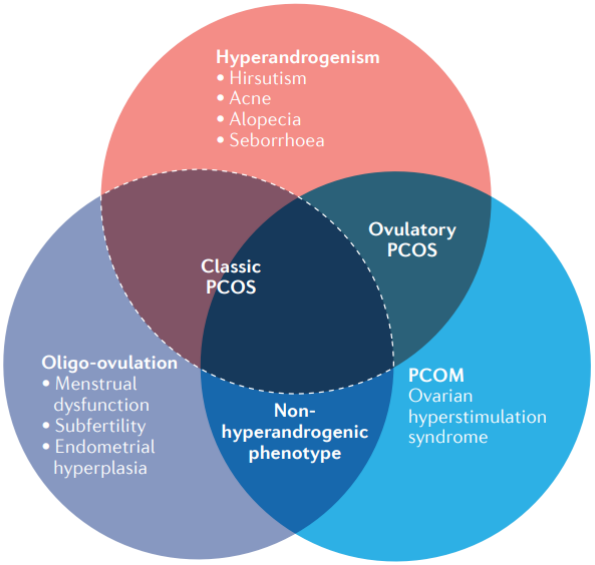

Figura 2.2: Subtipos del Síndrome de Ovario Poliquístico, en función de los síntomas [2].

el SOP puede provocar comorbilidades metabólicas y cardiovasculares, aunque los síntomas de hyperandrogenismo suelen disminuir durante la menopausia [8].

Las mujeres con SOP tienen mayor riesgo de sufrir diabetes tipo 2, infertilidad y complicaciones obstétricas, entre otras muchas patologías. A continuación se presenta la epidemiología del Síndrome de Ovario Poliquístico.

- Disfunción metabólica

  Muchas mujeres con SOP presentan hiperinsulinemia basal y estimulada por la glucosa y resistencia a la insulina, independientemente de su índice de masa corporal. La incidencia del síndrome metabólico, diabetes gestacional, intolerancia a la glucosa y diabetes tipo 2 es mayor en mujeres en premenopausia con SOP que en mujeres de edad e índice de masa corporal similar [9].

- Obesidad

  Hay una fuerte relación entre el SOP y la obesidad y viceversa, aunque no se ha podido demostrar que una sea la causa de la otra. Estadísticamente, la obesidad está más presente en mujeres con SOP que en mujeres sin SOP. No hay una evidencia clara de que la obesidad provoque el desarrollo o la prevalencia del SOP. Así mismo, no se ha encontrado relación entre variantes genéticas de la obesidad y genes relacionados con el SOP [10].

- Fertilidad

  El Síndrome de Ovario Poliquístico es la principal causa de infertilidad anovulatoria en mujeres. Según un estudio, el 72 % de las mujeres con SOP presentaban infertilidad, mientras que las mujeres sin SOP presentaban un 16 %, independientemente del índice de masa corporal. Estos porcentajes varían en otros estudios, aunque la tendencia se mantiene [11].

- Manifestaciones psicológicas

  Independientemente del fenotipo y de la presencia de obesidad, las mujeres con Síndrome de Ovario Poliquístico presentan más trastornos de depresión y ansiedad, y estos son más severos que en mujeres sin SOP. Algunos estudios señalan una relación entre el grado de depresión y de resistencia a la insulina [12].

- Problemas dermatológicos

  Los principales síntomas del hiperandrogenismo clínico incluyen hirsutismo (exceso de vello facial), acné y alopecia androgénica. El hirsutismo y el acné puede presentarse de forma distinta en función de la etnicidad de la paciente [13].

- Complicaciones obstétricas

  Las mujeres con SOP tienen mayor probabilidad de presentar complicaciones durante el embarazo. Estas complicaciones pueden abarcar partos prematuros, pre-eclapmsia y diabetes gestacional. También los niños nacidos pueden presentar síndrome de aspiración de meconio y bajo Apgar a los 5 minutos [14].

- Complicaciones cardiovasculares y cerebrovasculares

  Las mujeres con SOP tienen mayor prevalencia de marcadores de enfermedades cardiovasculares, como pueden ser la calcificación coronaria y aórtica. También la incidencia de accidentes cerebrovasculares es mayor en mujeres de edad avanzada con SOP. También presentan un mayor riesgo de sufrir tromboembolismo venoso [15].

- Riesgo de cáncer

  Las mujeres con SOP tienen un mayor riesgo de sufrir cáncer de endometrio. También se ha observado un mayor riesgo de cáncer ovárico, aunque el origen es incierto [16].

Capítulo 3

Metodología

3.1. Fundamentos

El Aprendizaje Automático o *Machine Learning* es un conjunto de métodos que detectan patrones en datos, y luego utiliza esos patrones para predecir datos a futuro, o para realizar otro tipo de toma de decisiones [17]. El Aprendizaje Automático se divide en tres tipos principales, en función de cómo se realiza el proceso mediante el que el algoritmo "aprende":

1. Aprendizaje supervisado: el propósito del modelo es obtener una variable objetivo y a través de la información de una serie de variables predictoras X. En este caso X puede ser un vector, o un objeto más complejo como una imagen, un texto, un grafo, etc. Igualmente, la variable objetivo y puede ser una variable categórica, en cuyo caso el algoritmo sería un algoritmo de clasificación o de reconocimiento de patrones; o una variable numérica, en cuyo caso se trataría de un problema de regresión.

2. Aprendizaje no supervisado: en este caso el algoritmo no tiene variable objetivo, solamente recibe como datos de entrada un conjunto de variables X, y el objetivo es aprender patrones de ese conjunto de variables. Un ejemplo de este tipo de aprendizaje sería el *clustering*.

3. Aprendizaje por refuerzo: el algoritmo aprende cómo desarrollarse en función de una señal de refuerzo o de penalización.

En este trabajo nos centraremos en los algoritmos de clasificación, ya que son los adecuados para tratar un problema de diagnóstico médico. A continuación se presentan algunos de los principales algoritmos de clasificación.

Decision tree

Un árbol de decisión clasifica los datos al resolver una serie de preguntas sobre las variables predictoras [18]. Cada una de estas preguntas se representa en una "hoja" o nodo, que divide los datos en dos grupos en función de la respuesta a esa pregunta. Estas preguntas pueden ser categóricas ("Sí" o "No") o numéricas, mediante un valor de corte en una variable predictora numérica. Esto se hace con cada una de las variables predictoras hasta llegar a una "jerarquía"de clasificación, que se representa en forma de árbol. Los árboles de decisión pueden ser más interpretables que otros clasificadores, ya que combinan una serie de preguntas sobre los datos que pueden ser entendidas fácilmente. De esta forma, pueden resultar muy útiles en sistemas de diagnóstico como el que pretendemos desarrollar, ya que se desarrolla un sistema de reglas muy similar al que se usa en el diagnóstico médico.

Los árboles de decisión se construyen tratando de separar los datos de entrenamiento de forma que se divida la variable objetivo lo más limpiamente posible. Para medir esta "limpieza" o calidad de la separación se pueden usar distintas métricas. Las más comunes son la entropía y

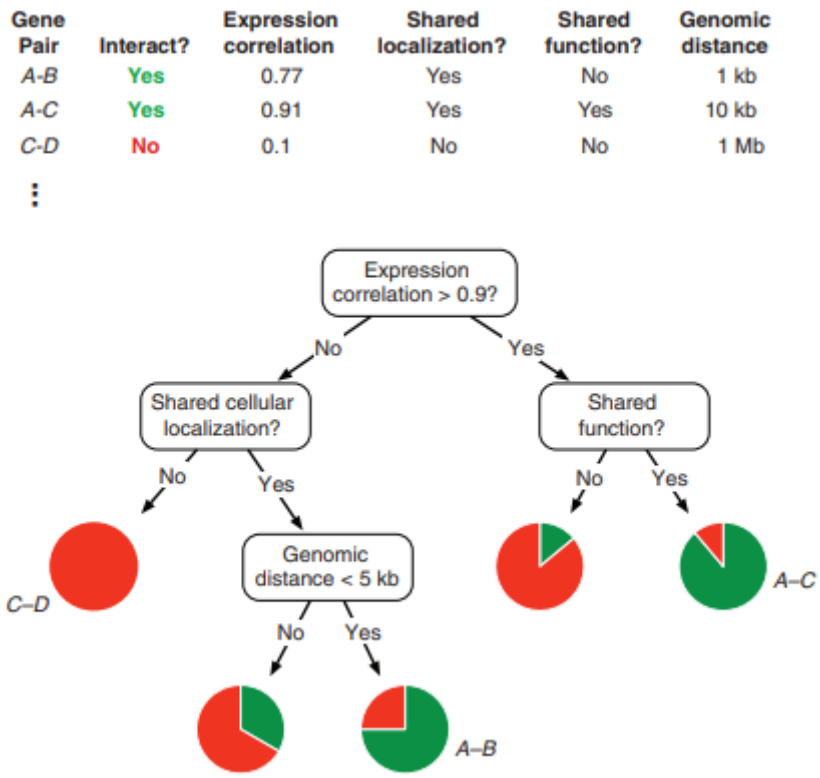

Figura 3.1: Construcción del árbol de decisión [18].

el índice Gini. La primera se refiere a la entropía de la distribución de probabilidad de los datos, que se describe [19]:

$$H = \sum_{i=1}^{n} p_i \log_2 p_i \tag{3.1}$$

Donde p_i es la probabilidad de cada distribución. Esta entropía es mínima cuando todos los elementos de un conjunto de datos pertenecen a una única clase de la variable objetivo, y es máxima cuando la distribución es equitativa, es decir, en nuestro caso cuando la mitad de los casos son positivos y la otra mitad negativos. El índice Gini funciona de manera similar, y también es mínimo cuando todos los elementos del conjunto pertenecen a la misma clase.

De esta forma, utilizando una métrica de la impureza de un conjunto de datos, se escogen las "preguntas" que minimizan esta impureza de los nodos resultantes. De esta forma se obtiene la ganancia de información o *information gain*, que representa la diferencia entre la impureza de los datos en el nodo padre y la impureza de los datos en los nodos hijos, dada una decisión. De esta forma se computan todas las decisiones posibles y se escoge la que produzca mayor ganancia de información.

Una vez construido el árbol de decisión, es importante saber hasta qué punto dividir los nodos del árbol, para tratar de evitar el sobre-ajuste de los datos o *overfitting*. Esto se puede hacer determinando un número máximo de nodos del árbol, o seleccionando un umbral de la ganancia de información: cuándo añadir un nodo aumente la ganancia por debajo de un umbral, se deja de crecer el árbol. Otra forma de evitar el sobre-ajuste es dejar que el árbol crezca todo lo posible, y después eliminar nodos, en función del error en la clasificación que produzcan.

Random Forest

El algoritmo *Random Forest* consiste en la creación de una serie de árboles de decisión, que más tarde se agregan para formar un sistema de clasificación mejor que el que podría formar un único árbol [20]. En primer lugar, se divide el conjunto de entrenamiento en una serie de muestras aleatorias, cada una de las cuales se utiliza para generar un árbol de decisión. Estos árboles de decisión se entrenan hasta alcanzar el menor error en la clasificación posible. Este conjunto de árboles forma un "bosque" a partir del cual se realiza la clasificación, combinando las predicciones de cada árbol, atendiendo a la clase mayoritaria.

Para determinar la eficacia de estos bosques aleatorios, hay que decidir una serie de parámetros, como el número de árboles que se producen, la profundidad de esos árboles, la forma en la que se escogen los conjuntos de datos, etc.

Gradient Boosting y *Extreme Gradient Boosting*

En *Gradient Boosting*, al contrario que en el caso anterior, los árboles de decisión se construyen secuencialmente, para minimizar el error del árbol anterior [21]. Esto puede resultar en un gran coste de computación, por lo que surgió el algoritmo *Extreme Gradient Boosting*. Este algoritmo reduce el tiempo de computación al organizar los datos para reducir el tiempo de búsqueda, y utilizar computación en paralelo.

Adaptive Boosting

Este algoritmo también utiliza *boosting*, pero funciona de forma diferente a los anteriores. En este caso se construyen una serie de clasificadores secuencialmente, de forma que cada clasificador se centra en los casos del conjunto de entrenamiento que han sido mal clasificados por los clasificadores anteriores [22]. Una vez producidos estos clasificadores, se escogen aquellos que produzcan la mejor clasificación con un conjunto de *test*.

Support Vector Machine

Este algoritmo trata de clasificar los datos de entrenamiento encontrando una superficie en el espacio de las variables que separe lo máximo posible la variable objetivo [23]. Si partimos de un ejemplo en dos dimensiones, donde queremos separar dos grupos de puntos, el objetivo del *SVM* es encontrar el hiperplano que mejor separe estos dos conjuntos de puntos, dentro de una colección de posibles hiperplanos.

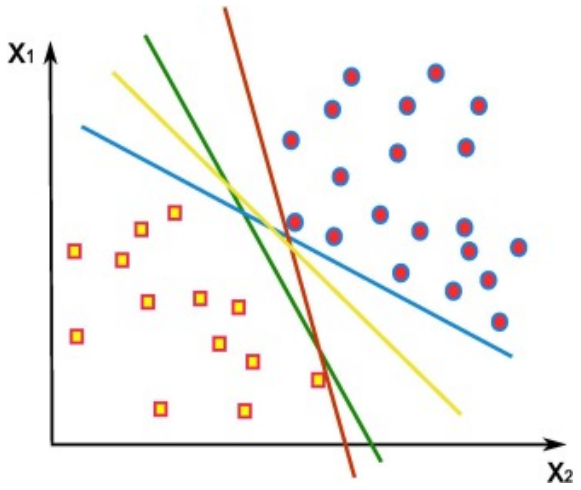

Figura 3.2: Ejemplo de *clustering* de dos grupos de puntos utilizando *SVM* [23].

En la Figura 3.2 se pueden observar varios hiperplanos (que en este caso al utilizar datos en dos dimensiones se representan como rectas) que podrían separar los dos conjuntos de datos. De esta forma, el objetivo del modelo es encontrar cuál de esas rectas separa mejor los puntos, es decir la recta que produzca un mayor margen entre los dos grupos de puntos y la recta. Esto se hace calculando la distancia entre las rectas y los puntos de los datos, y esta distancia se maximiza. A los puntos que están más cerca de estas rectas se les denomina *support vectors*, ya que son los que más participan en la definición del hiperplano, de forma que el resto de puntos pueden ser irrelevantes para el modelo.

En este ejemplo hemos utilizado dos conjuntos de datos que son linealmente separables, aunque este no es siempre el caso de todos los modelos que vayamos a desarrollar. Cuando los datos no son linealmente separables, puede ser necesario transformar el espacio de las variables de entrada a un espacio con mayor dimensionalidad, de forma que en este nuevo espacio, denominado "espacio de características dimensionales" es más fácil encontrar un hiperplano que separe los datos. Esto se hace a través de una función de *kernel*, que transforma el espacio de las variables de entrada en un espacio distinto. Existen muchas funciones de kernel, aunque aquí exponemos las más comunes:

- Kernel lineal: $K(x_i, x_j) = (x_i \cdot x_j)$
- Kernel polinómico: $K(x_i, x_j) = (x_i \cdot x_j + 1)^p$
- Kernel gaussiano: $K(x_i, x_j) = e^{\frac{-||x_i - x_j||^2}{2\sigma^2}}$
- Kernel RBF: $K(x_i, x_j) = e^{-\gamma(x_i - x_j)^2}$
- Kernel sigmoide: $K(x_i, x_j) = \tanh(\eta x_i \cdot x_j + \nu)$

Neural network

Las redes neuronales o *neural network* es un algoritmo que se basa en el funcionamiento del cerebro a la hora de procesar y almacenar información. Utiliza un gran número de nodos o "neuronas" que se conectan entre ellos. Cada nodo representa una función con entrada y salida, llamada función de activación. Estas neuronas se conectan entre ellas a través de "pesos", que guardan la información de cada operación [24]. Estas neuronas se agrupan por capas, de forma que al añadir muchas capas internas se produce lo que se conoce como Aprendizaje Profundo o *Deep Learning*. El resultado de la red neuronal dependerá del número de neuronas, de capas y del tipo de función de activación que se utiliza en cada capa.

En la Figura 3.3 se puede ver un esquema de la estructura de una red neuronal. En este caso se puede observar que la capa de entrada tiene 3 neuronas, en cada una de las cuales se aplica una función de activación, que transformará los datos de entrada. Si esta función de activación es una función lineal, cada neurona transformará los datos de entrada añadiendo un peso y un factor de sesgo, de forma que el resultado de esa neurona sea una función lineal $y = wx + b$. Estos valores pasan a las neuronas de la siguiente capa oculta, en este caso al haber 3 neuronas en la capa de entrada, y 4 neuronas en la capa oculta, cada neurona de la capa oculta recibe tres funciones distintas. Utilizando los pesos y sesgos que recibe de la capa de entrada, se vuelven a transformar siguiendo la función de activación de la capa oculta. Este proceso se repite en cada capa hasta llegar a la capa final, donde se produce la predicción de la red neuronal.

Este es un ejemplo de una red neuronal sencilla, pero hay más tipos de redes neuronales: redes prealimentadas, recursivas, convolucionales, etc. Cada uno de estos tipos tiene aplicaciones distintas, y son utilizadas en una gran cantidad de campos.

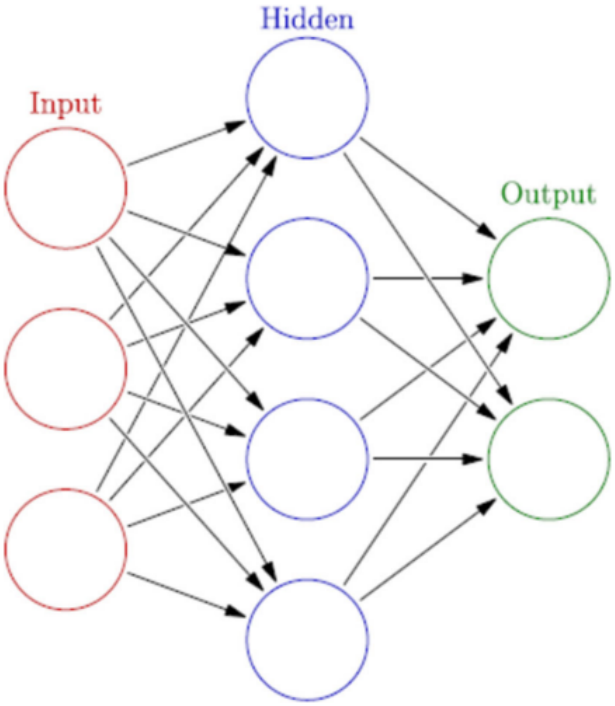

Figura 3.3: Estructura de una red neuronal con una capa oculta [24].

Stacking

La combinación de modelos consiste en combinar una serie de modelos de Aprendizaje Automático para producir otro modelo más fuerte [25]. Estos modelos se pueden combinar de dos formas: *bagging*, donde se combinan los modelos aplicando un peso, para compensar las distintas características que puedan tener cada modelo; y *Stacking*, donde se utilizan las predicciones de los modelos a combinar como datos de entrenamiento para el siguiente modelo.

Estos son solamente algunos de los principales modelos de clasificación de Aprendizaje Automático. Existen muchos más: *KNN*, redes bayesianas, *LDA*, etc. En este trabajo nos vamos a centrar en aquellos algoritmos que son los más adecuados para tratar un problema de diagnóstico médico. Hemos decidido seleccionar tres algoritmos: *Decision tree*, que es un algoritmo simple pero fácilmente interpretable, muy adecuado para problemas de diagnóstico; *Random Forest*, que se basa en árboles de decisión pero es más potente y puede producir mejores resultados; y *Support Vector Machine*, que es un algoritmo más general y que funciona mejor con pocos datos. Así mismo, para comprobar si estos algoritmos eran los más adecuados o no, se han hecho pruebas utilizando otros algoritmos, que se pueden observar en el Apéndice A.

3.2. Estado del arte

Se ha encontrado un estudio que utiliza Inteligencia Artificial para diagnosticar el Síndrome de Ovario Poliquístico [26]. Este estudio trata de desarrollar un modelo que asegure la eficiencia en el diagnóstico del SOP. Para ello utiliza varios algoritmos de Aprendizaje Automático, como regresión logística, árboles de decisión, redes bayesianas y otros. También se propuso un modelo que combinaba las predicciones de todos los modelos anteriores para mejorar el funcionamiento del modelo de diagnóstico, junto con otras técnicas de optimización. En este estudio se alcanzó una precisión o accuracy del 98,87 %.

3.3. Aplicaciones de IA en medicina

En los últimos años se ha utilizado la Inteligencia Artificial como herramienta de investigación en muchos ambientes médicos. Por ejemplo, la Inteligencia Artificial se ha aplicado en el campo de la anestesiología, en particular en el estudio de la supervisión y el control de la anestesia, prevención de riesgos, guía de ultrasonidos, el manejo del dolor y la logística quirúrgica [28]. Así mismo, el diseño farmacológico se ha visto beneficiado por esta técnica. El diseño de fármacos es una parte muy importante de la investigación farmacéutica. Este proceso se enfrenta a muchos problemas que pueden ser solucionados usando Inteligencia Artificial, como la síntesis de péptidos, el cribado virtual basado en estructuras, el cribado virtual basado en ligandos, la predicción de toxicidad, la monitorización y liberación de fármacos, etc [27].

El Aprendizaje Automático y Profundo puede ser una herramienta muy útil para la cardiología, ya que generalmente esta disciplina trabaja con grandes cantidades de información sobre los pacientes [29]. Una parte importante de esta especialización es tratar de hacer predicciones sobre la salud cardiovascular de un paciente en función de grandes cantidades de información sobre su salud general. Es aquí donde la Inteligencia Artificial puede suponer un gran avance. Ya se han realizado estudios sobre los riesgos de accidentes cardiovasculares, la construcción de modelos predictivos para infarto agudo de miocardio usando medidas proteómicas y variables clínicas, la predicción de la reestenosis de stent a partir de metabolitos plasmáticos, etc.

Igualmente, desde hace años se utiliza el Aprendizaje Profundo para muchas partes del estudio de la patología tumoral: el diagnóstico de tumores, identificación de subtipos de tumores, graduación, determinación del estado del tumor, predicción de pronóstico e identificación de características patológicas, biomarcadores y cambios genéticos [30]. Este campo se ha visto enormemente beneficiado por el uso de estas herramientas, que alivian la carga de trabajo de los científicos en muchas partes vitales del tratamiento oncológico.

De igual forma, el Aprendizaje Profundo es excepcionalmente bueno en la tarea de análisis de imágenes. Es debido a esto que se ha descubierto como una herramienta útil para el análisis de imágenes médicas [31]. Por ejemplo, la tarea de determinar el borde de un tumor cancerígeno debía ser realizada por un personal sanitario que observase cada imagen y tratase de determinar dónde se producía esta división entre tejido cancerígeno y tejido sano. Actualmente, se han desarrollado modelos de Aprendizaje Profundo capaces de realizar esta tarea de forma automática. Así mismo, el Aprendizaje Automático y la Visión Artificial se han convertido en herramientas muy importantes en la microcirugía oftalmológica [32]. La Inteligencia Artificial ya se ha aplicado en este campo a la detección retinopatías prematuras, retinopatías diabéticas, degeneración macular asociada con la edad y glaucoma.

3.4. Ética y consideraciones legales del uso de IA en Medicina

Existe un debate en la sociedad sobre el uso y alcance de la Inteligencia Artificial, y este debate se ve ampliado cuando se aplica a la medicina. La ética de la Inteligencia Artificial se puede ver como una ampliación de la ética aplicada a la era digital, que analiza los problemas que surgen por el uso y avance de todas las tecnologías digitales, entre ellas la Inteligencia Artificial, el *Big Data* y el Aprendizaje Automático [33]. En los últimos años ha surgido una preocupación respecto al uso de la Inteligencia Artificial, tanto por el mal uso de esta tecnología como por los errores en el diseño de algunas de sus aplicaciones. Debido a esto algunas industrias han visto la necesidad de crear unas guías de uso de esta nueva tecnología, como los Principios de Inteligencia Artificial de Google [34]. Estos principios suelen contener una serie de valores que respetar a la hora de utilizar o desarrollar Inteligencia Artificial. El principal problema de estos principios es que no están unificados, de forma que se pueden producir incongruencias entre algunos de ellos; y muchos pueden resultar poco específicos.

Así mismo, ha habido algunos intentos de legislar el uso y desarrollo de la Inteligencia

Artificial. En la Unión Europea se aprobó recientemente un Reglamento de Inteligencia Artificial que regulará el desarrollo de la Inteligencia Artificial para garantizar que esta tecnología sea segura y respete los derechos de los ciudadanos [35]. Esta regulación estará basada en un sistema de riesgos, que clasificará cada sistema de Inteligencia Artificial en una categoría en función del riesgo que pueda suponer para los ciudadanos, y en función de la categoría a la que pertenezca se legislará de una forma distinta su desarrollo y aplicación. En este reglamento, los sistemas de Inteligencia Artificial que se utilizan para tomar decisiones médicas se clasificarían en alto riesgo.

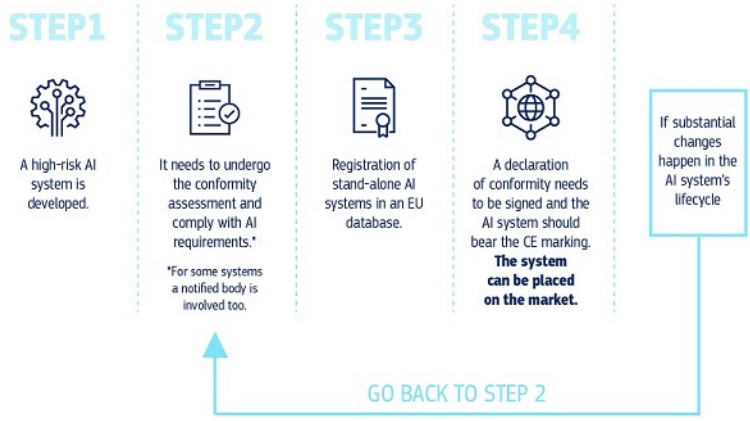

Figura 3.4: Funcionamiento del Reglamento de Inteligencia Artificial de la Unión Europea [35].

Las tecnologías que estén clasificadas como alto riesgo deberán someterse a una estricta regulación, que incluye introducir sistemas de gestión de riesgos, asegurar la trazabilidad de los resultados, documentación detallada, etc.

Específicamente en el campo de la medicina, muchos profesionales se han preocupado por el uso de la Inteligencia Artificial, ya que si se utiliza inadecuadamente puede agravar desigualdades existentes en el sector de la salud [36]. Algunos de estos profesionales se preocupan por si el uso de la Inteligencia Artificial puede respetar la privacidad de los pacientes, al mismo tiempo que agravar las desigualdades de género y raza. Debido a esto se deben desarrollar normativas específicas para el uso de la Inteligencia Artificial en la medicina, que aseguren el desarrollo de esta tecnología de una forma ética y moralmente apropiada, de forma que se beneficie a toda la sociedad.

Capítulo 4

Dataset

4.1. Datos utilizados

Para realizar este trabajo se ha utilizado dos conjuntos de datos. El primero contiene información de 541 pacientes [37]. Este *dataset* contiene un total de 42 variables, entre las cuales se incluye el diagnóstico positivo o negativo de SOP (variable objetivo). Así mismo, se ha utilizado un segundo conjunto de datos [38], que incluye información de 465 pacientes. Este *dataset* abarca 16 variables, entre las que se encuentra si la paciente tiene un diagnóstico previo de SOP o no. Ambos *datasets* se encontraron en la página *Kaggle* y son de libre acceso. En la Tabla 4.1 se presentan las variables de ambos *datasets*.

Dataset	**Variable**	**Descripción**	**Tipo de variable**	**Unidad de medida**
1 y 2	Edad	Edad de la paciente	Numérica	años
1 y 2	Peso	Peso de la paciente	Numérica	Kg
1 y 2	Altura	Altura de la paciente	Numérica	cm
1	BMI	Índice de masa corporal, que se calcula a partir del peso y altura de la paciente: IMC=peso(kg)/[altura(m)]ˆ2	Numérica	kg/mˆ2
1 y 2	Grupo sanguíneo	Se transforma en un número del 11 al 18 en función del grupo sanguíneo (A+, A-, B+, B-, O+, O-, AB+, AB-)	Numérica	
1	Frecuencia cardiaca	Frecuencia cardiaca de la paciente	Numérica	puls/min
1	Respiraciones por minuto	Respiraciones por minuto de la paciente	Numérica	resp/min
1	Hb	Hemoglobina	Numérica	g/dL
1 y 2	Ciclo	Indica si el ciclo es regular o irregular	Categórica	
1 y 2	Duración del ciclo	Duración de la menstruación en días	Numérica	días
1	Estado civil	Indica si la paciente está casada o no	Categórica	
1	Embarazo	Indica si la paciente está embarazada o no	Categórica	
1	Número de abortos	Número de abortos que ha sufrido la paciente	Numérica	

1	I beta-HCG	Hormona beta-HCG	Numérica	mlU/mL
1	II beta-HCG	Segunda prueba de beta-HCG (en las pacientes embarazadas se realiza unos días después de la primera, si no se realiza se pone el resultado de la primera prueba)	Numérica	mlU/mL
1	FSH	Hormona foliculoestimulante	Numérica	mlU/mL
1	LH	Hormone luteinizante	Numérica	mlU/mL
1	FSH/LH	Proporción entre la hormona FSH y la LH	Numérica	
1	Cadera	Contorno de cadera	Numérica	pulgadas
1	Cintura	Contorno de cintura	Numérica	pulgadas
1	Cadera/Cintura	Proporción entre el contorno de cadera y el contorno de cintura	Numérica	
1	TSH	Tirotropina	Numérica	mlU/mL
1	AMH	Hormona anti-mülleriana	Numérica	ng/mL
1	PRL	Prolactina	Numérica	ng/mL
1	Vit D3	Vitamina D3	Numérica	ng/mL
1	PRG	Progesterona	Numérica	ng/mL
1	RBS	Prueba aleatoria de glucosa en sangre	Numérica	mg/dL
1 y 2	Aumento de peso	Indica si la paciente ha sufrido aumento de peso o no	Categórica	
1 y 2	Crecimiento de pelo	Indica si la paciente ha sufrido crecimiento de pelo de patrón andrógino o no	Categórica	
1 y 2	Oscurecimiento de piel	Indica si la paciente ha sufrido oscurecimiento de piel o no	Categórica	
1 y 2	Caída de pelo	Indica si la paciente ha sufrido caída de pelo o no	Categórica	
1 y 2	Acné	Indica si la paciente ha sufrido acné o no	Categórica	
1 y 2	Comida rápida	Indica si la paciente consume comida rápida habitualmente o no	Categórica	
1 y 2	Ejercicio regular	Indica si la paciente realiza ejercicio regularmente o no	Categórica	

1	Presión sistólica	Presión sistólica de la paciente	Numérica	mmHg
1	Presión diastólica	Presión diastólica de la paciente	Numérica	mmHg
1	Número de folículos izquierdo	Número de folículos en el ovario izquierdo	Numérica	
1	Número de folículos derecho	Número de folículos en el ovario derecho	Numérica	
1	Tamaño medio folículos izquierdo	Tamaño medio de los folículos del ovario izquierdo	Numérica	mm
1	Tamaño medio folículos derecho	Tamaño medio de los folículos del ovario derecho	Numérica	mm
1	Endometrio	Grosor del endometrio	Numérica	mm
2	Meses periodo	Indica cada cuántos meses la paciente tiene el periodo	Numérica	
2	Cambios de humor	Indica si la paciente ha sufrido cambios de humor o no	Categórica	

Tabla 4.1: Variables de ambos *datasets*.

Como se puede observar, ambos *datasets* coinciden en muchas variables, pero el primero es mucho más completo, ya que cuenta con resultados de pruebas médicas, mientras el segundo solo cuenta con síntomas y datos básicos que se pueden recoger en una consulta. Así mismo, el segundo *dataset* no incluye la variable del índice de masa corporal, aunque sí incluye las variables necesarias para obtenerlo, la altura y el peso.

4.2. Primer *dataset*

En cuanto al primer *dataset*, en la Figura 4.1 se pueden observar las correlaciones entre la variable objetivo y el resto de variables, obtenidas mediante el método Pearson. Como se puede observar, las variables que tienen más correlación con la variable objetivo son aquellas que según el conocimiento experto son más utilizadas para diagnosticar el SOP: la presencia de ovario poliquístico (número de folículos por ovario) y los síntomas de hiperandrogenismo clínico (oscurecimiento de piel y crecimiento de pelo de patrón andrógino). De esta manera podemos comprobar de forma sencilla que los criterios que se utilizaron para diagnosticar a estas pacientes coinciden con la información que poseemos sobre el SOP.

En la Figura 4.2 se puede observar un mapa de calor de la correlación de todas las variables del primer *dataset*. De este modo quedan resaltadas las correlaciones más significativas, como por ejemplo las correspondientes a las variables de proporción: proporción cintura/cadera y proporción FSH/LH. También destaca una relación similar entre las variables peso, altura y BMI. De forma menos evidente se observa un pequeño grupo de variables con una correlación media donde se encuentran las variables correspondientes a los síntomas de hiperandrogenismo

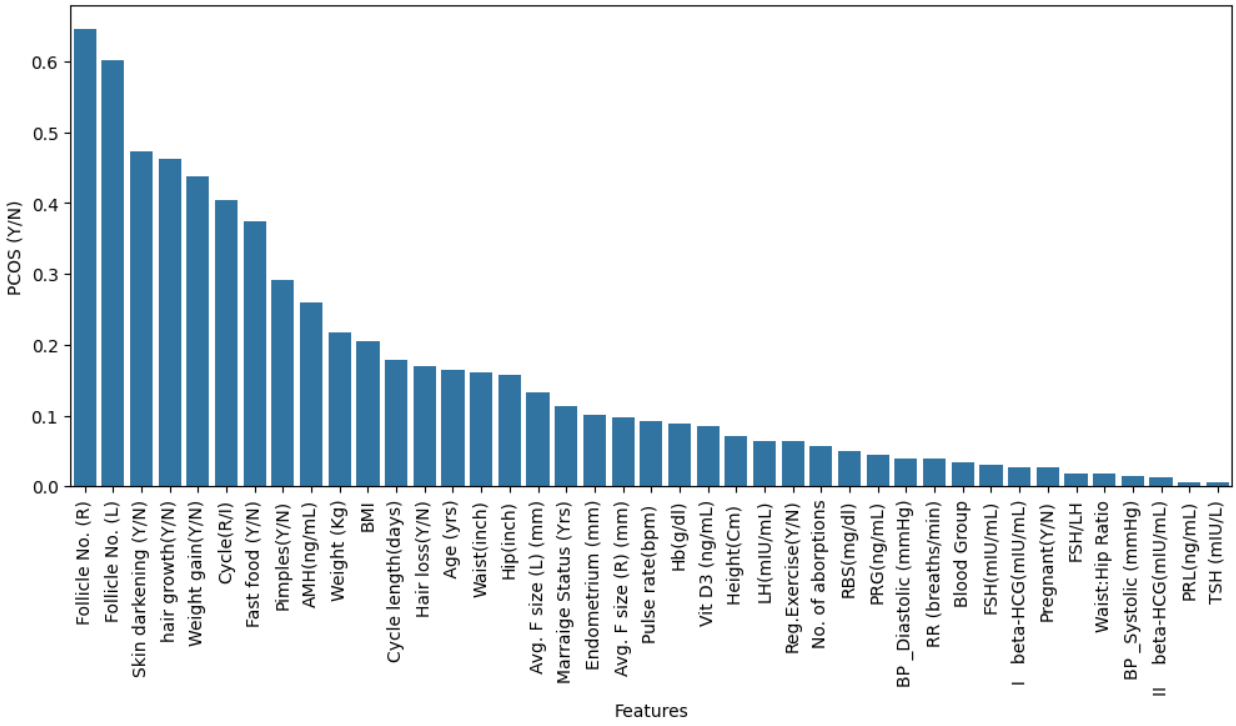

Figura 4.1: Correlación entre la variable objetivo y el resto de variables del primer *dataset*. En el eje x se encuentran las variables ordenadas de mayor a menor correlación.

clínico. Esto mismo pasa con otras variables que están relacionadas: altura y peso, cintura y cadera, número de folículos en el ovario izquierdo y derecho, etc.

En la Figura 4.3 se puede observar la distribución de casos positivos y negativos de SOP por grupos de edad. En este caso se puede observar que el primer *dataset* no contiene casos de menores de edad, y la mayoría de las pacientes se encuentran en el rango de 26 a 35 años. Esto coincide con algunos estudios, que indican la edad media de diagnóstico del Síndrome de Ovario Poliquístico en torno a los 26 años [39].

En la Figura 4.4 se puede observar el histograma del índice de masa corporal (BMI por sus siglas en inglés) separado en positivos y negativos al diagnóstico de SOP. Como se puede observar, en ambos casos el IMC parece seguir una distribución normal. Para comprobar la normalidad de estas distribuciones, se realizó un test de Kolmogórov-Smirnov (una prueba no paramétrica que compara dos distribuciones de probabilidad e indica cuánto se parecen), comparándolas con una distribución normal de misma media y desviación estándar, obteniendo los siguientes resultados:

Distribución	Statistic	p-value
Casos positivos	0,057	0,683
Casos negativos	0,036	0,862

Tabla 4.2: Resultados del test de Kolmogórov-Smirnov de las distribuciones del índice de masa corporal de casos positivos y negativos de SOP del primer *dataset*.

De esta forma, utilizando un valor de corte de $\alpha = 0{,}05$, no podemos descartar la hipótesis de que ambas distribuciones pertenecen a una normal, aunque se puede determinar que la distribución correspondiente a los casos negativos se parece más a una normal que la correspondiente a los casos positivos.

En la Figura 4.5 se puede observar el histograma del número de folículos en cada ovario separado en positivos y negativos al diagnóstico de SOP.

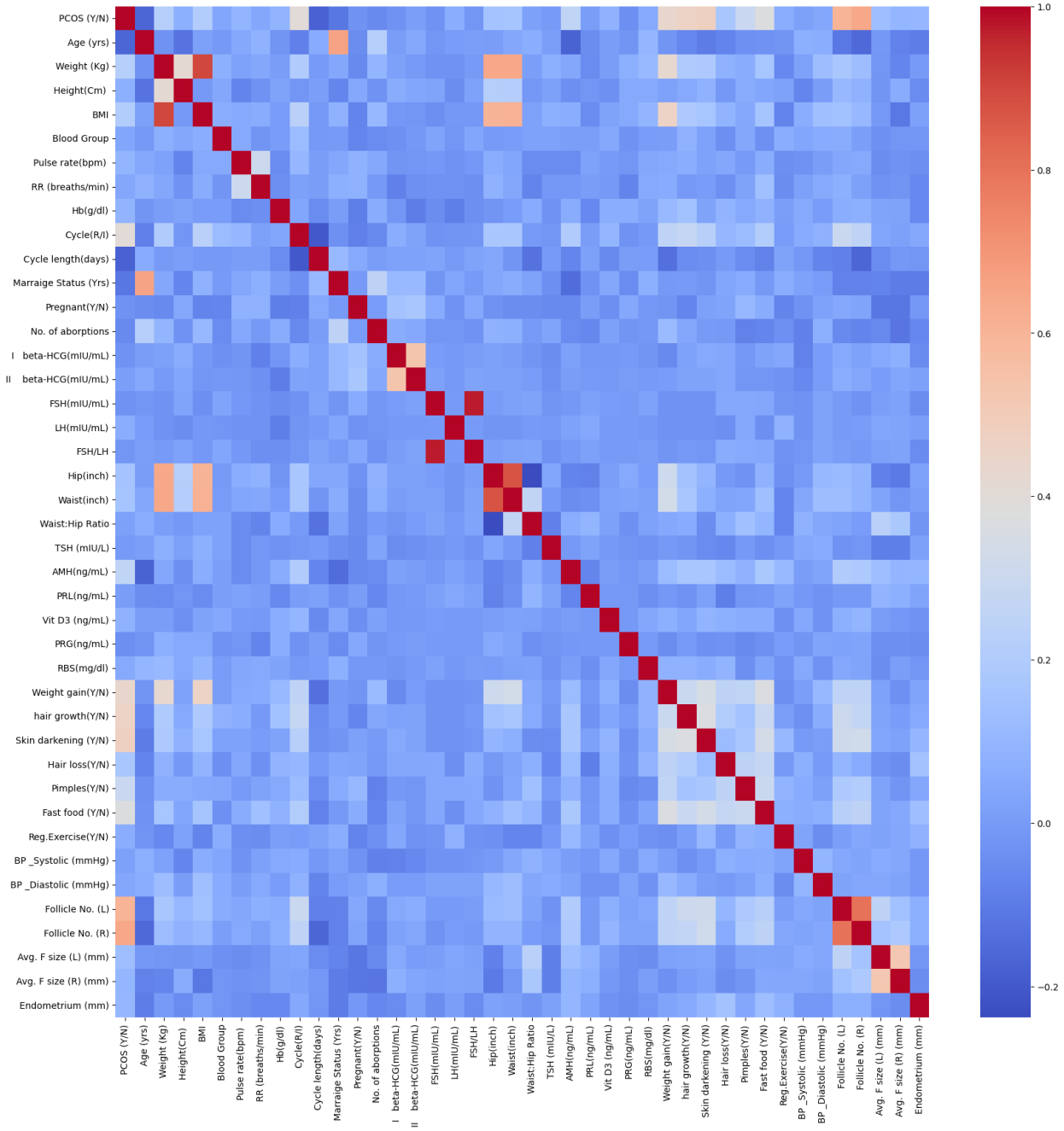

Figura 4.2: Mapa de calor de las variables del primer *dataset*. Se muestran las correlaciones de cada variable con el resto de variables, de color azul a rojo en función, de su correlación.

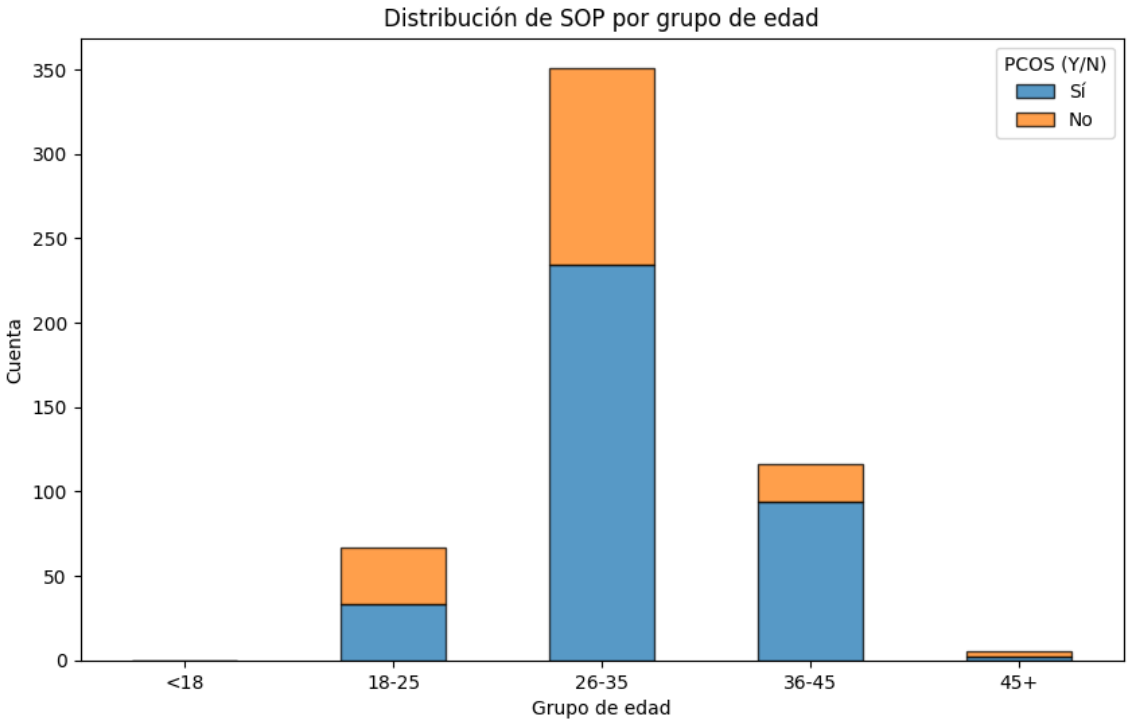

Figura 4.3: Distribución de positivos (azul) y negativos (naranja) de SOP por grupos de edad, para el primer *dataset*.

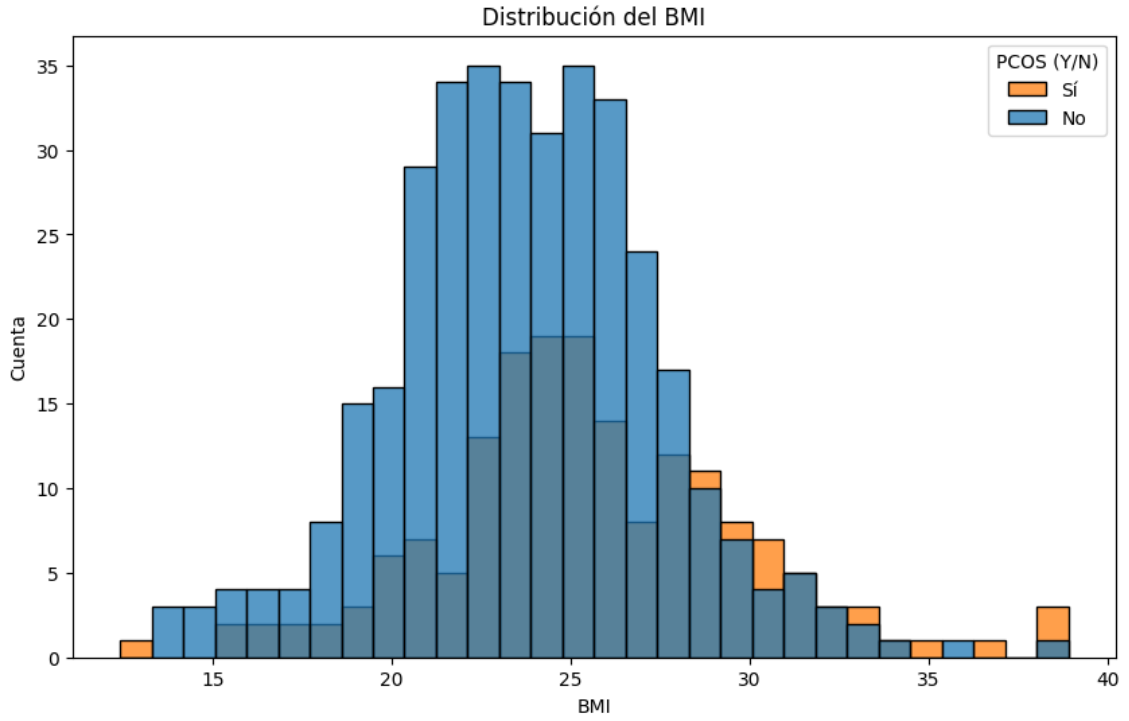

Figura 4.4: Distribución del índice de masa corporal del primer *dataset* de positivos (naranja) y negativos (azul) de SOP.

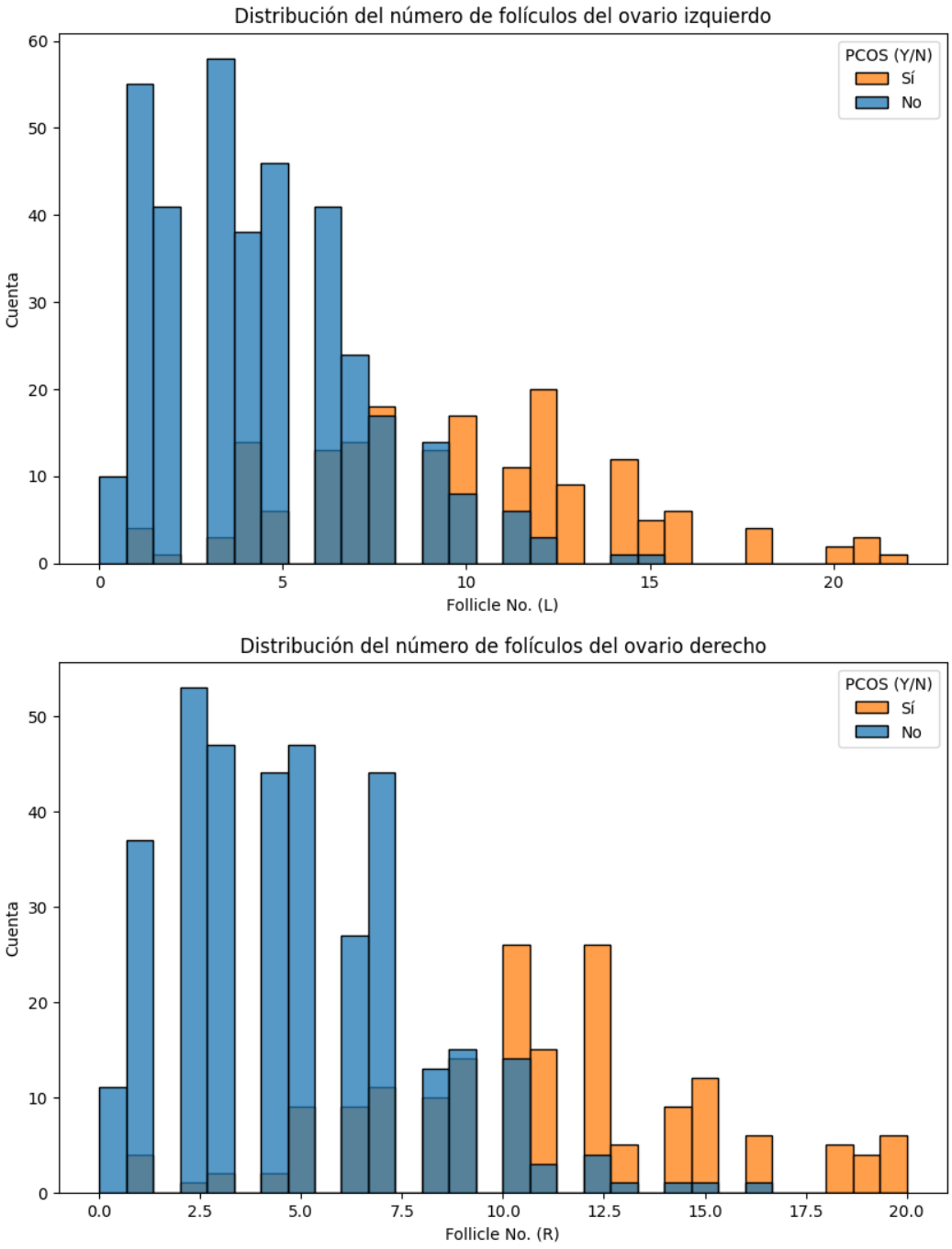

Figura 4.5: Distribución del número de folículos en ovario izquierdo (arriba) y ovario derecho (abajo) del primer *dataset*, separado por positivos (naranja) y negativos (azul) de SOP.

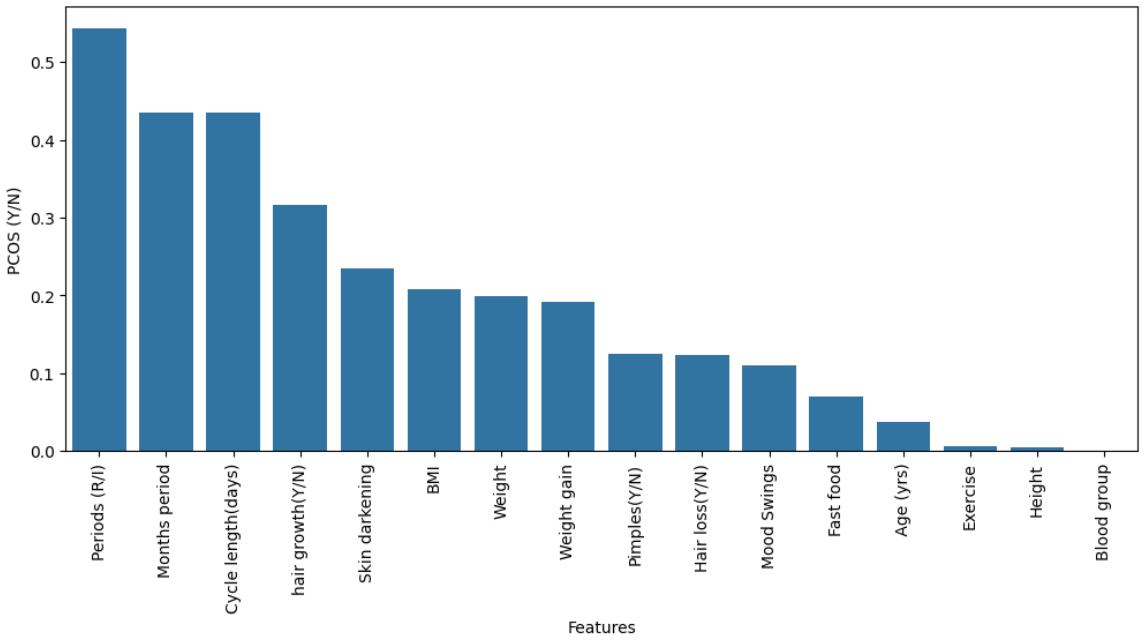

Figura 4.6: Correlación entre la variable objetivo y el resto de variables del segundo *dataset*. En el eje X se encuentran las variables ordenadas de mayor a menor correlación.

Al igual que en los histogramas anteriores, el número de datos con diagnóstico negativo es mayor que el de diagnóstico positivo. Igualmente, se observa como en los casos positivos el número de folículos es mayor que en los casos negativos, lo cual concuerda con el criterio de diagnóstico del conocimiento experto.

De esta forma se concluye que el primer *dataset* utilizado en este trabajo es heterogéneo, ya que hay más casos de diagnóstico negativo que positivo, y tanto la correlación entre las variables como su distribución concuerdan con los criterios de diagnóstico del SOP actuales.

4.3. Segundo *dataset*

En cuanto al segundo *dataset* utilizado, en la Figura 4.6 se puede observar la correlación de cada variable con la variable objetivo, obtenida mediante el método Pearson. Cabe destacar que para comparar este *dataset* con el primero que se ha expuesto, y para resultar más útil a la hora de realizar los modelos de diagnóstico, se ha creado la variable *BMI*, correspondiente al índice de masa corporal, a partir de los valores de las variables *Height* (altura) y *Weight* (peso).

Como se puede observar, al igual que en el primer *dataset*, las variables que tienen mayor correlación con la variable objetivo son aquellas que se utilizan en la actualidad para diagnosticar el SOP, principalmente los síntomas de hiperandrogenismo clínico y la irregularidad del ciclo menstrual.

En la Figura 4.7 se puede observar el mapa de calor que representa las correlaciones entre todas las variables. En este caso se puede observar una gran correlación entre la duración del ciclo (*Cycle length(days)*) y cada cuántos meses tiene la paciente el periodo (*Months period*), así como entre el índice de masa corporal y el peso, si bien esta correlación es menor en el caso del índice de masa corporal y la altura. También se observa, al igual que en el primer *dataset*, un grupo de variables con una ligera correlación, donde se encuentran las variables que indican los síntomas del hiperandrogenismo clínico.

En la Figura 4.8 se presenta la distribución de positivos y negativos de SOP del segundo *dataset*, dividida en grupos de edad. En este caso se puede observar la diferencia más significativa entre el primer *dataset* y el segundo: el último contiene datos de pacientes menores de edad.

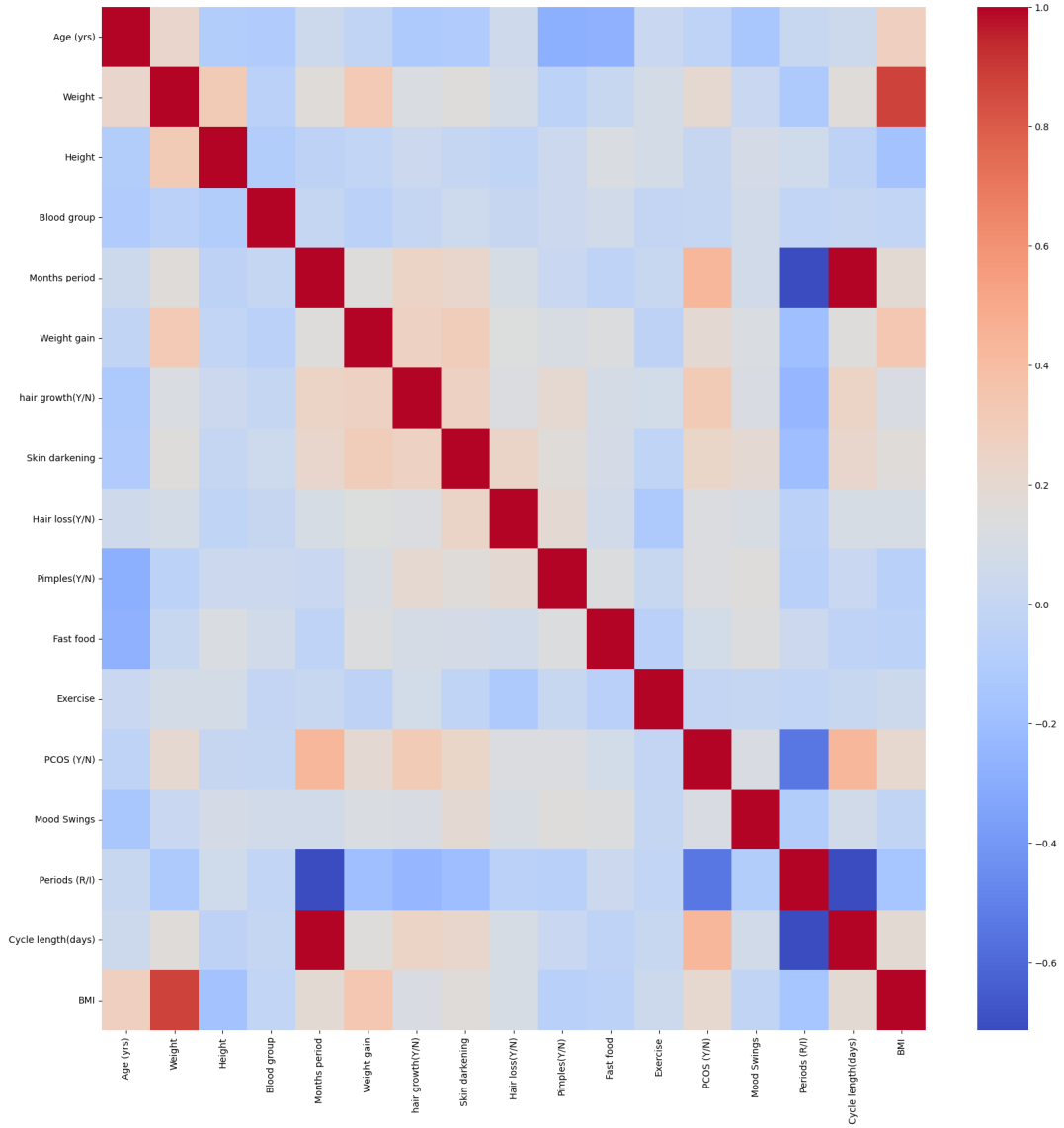

Figura 4.7: Mapa de calor de las variables del segundo *dataset*. Se muestran las correlaciones de cada variable con el resto de variables, con un color de azul a rojo en función de su correlación.

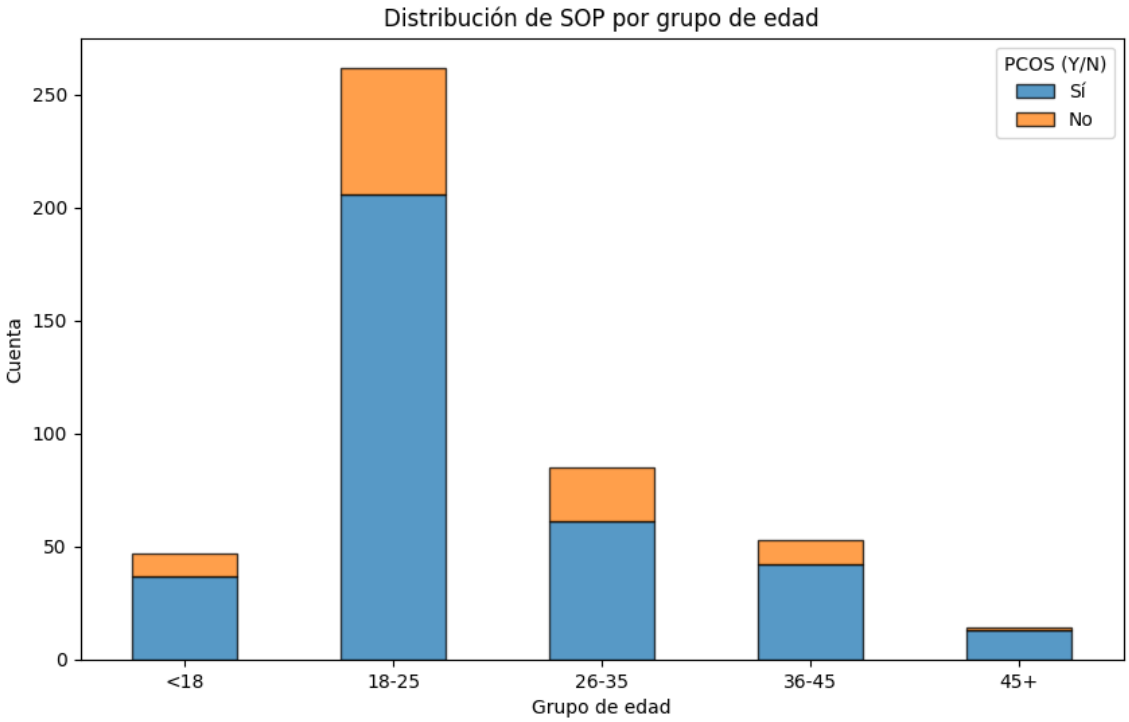

Figura 4.8: Distribución de positivos (azul) y negativos (naranja) de SOP del segundo *dataset* por grupos de edad.

Además, la mayoría de los datos de este *dataset* se encuentran en el grupo de edad de 18 a 25 años, mientras que en el primero se encontraba en el siguiente grupo, de 26 a 35 años. Esto puede resultar llamativo, aunque el pico de diagnósticos se encuentra cerca de la edad media de diagnóstico, en torno a los 26 años.

En la Figura 4.9 se presenta la distribución del índice de masa corporal (*BMI* por sus siglas en inglés) del segundo *dataset* dividida en casos positivos y negativos de SOP. En este caso se puede observar, al igual que en el primer *dataset*, que el número de casos negativos es claramente mayor que el de casos positivos. Así mismo, se observa que ambos parecen seguir distribuciones normales, por lo que se realizó un test de Kolmogórov-Smirnov para comprobar la normalidad de estas distribuciones, obteniendo los siguientes resultados:

Distribución	Statistic	p-value
Casos positivos	0,076	0,635
Casos negativos	0,063	0,228

Tabla 4.3: Resultados del test de Kolmogórov-Smirnov de las distribuciones de índice de masa corporal de casos positivos y negativos de SOP del segundo *dataset*.

En este caso, al igual que en el primer *dataset*, con un valor de corte de $\alpha = 0{,}05$ en ninguno de los casos se puede descartar la hipótesis de que ambas distribuciones procedan de una distribución normal. Sin embargo, al contrario que en el caso anterior, en este *dataset* la distribución correspondiente a los casos positivos se parece más a una distribución normal que la correspondiente a los casos negativos.

A primera vista estos *datasets* son obviamente distintos: el primero cuenta con 42 variables, y el segundo solamente con 16. El primer *dataset* incluye resultados de pruebas médicas, principalmente la ecografía ovárica y analítica sanguínea; mientras que el segundo solamente incluye los síntomas de las pacientes, y algún dato básico como el peso y altura, el grupo sanguíneo y

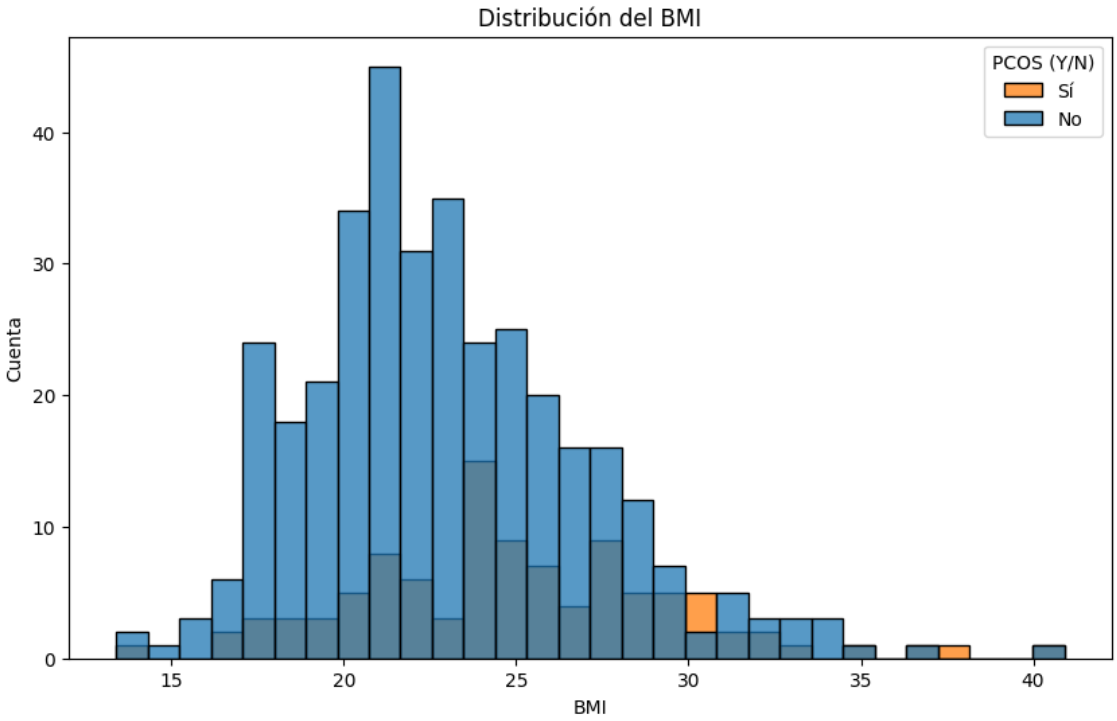

Figura 4.9: Distribución del índice de masa corporal del segundo *dataset*, dividida en positivos (naranja) y negativos (azul).

sobre el estilo de vida de la paciente. Esto en un primer momento puede parecer un problema, pero realmente podemos darle un uso a ambos *datasets*, ya que parte del objetivo de este trabajo es crear un modelo de diagnóstico que sea capaz de predecir si una paciente tiene SOP sin la necesidad de realizar pruebas médicas, es decir, utilizando solamente variables básicas como las que encontramos en el segundo *dataset*.

Para estudiar más a fondo las diferencias entre los dos *datasets* presentados, en primer lugar se observaron los histogramas de las variables más significativas de ambos. En la Figura 4.10 se representan el número de diagnósticos positivos y negativos de cada *dataset*. Como se puede observar, es de especial interés destacar que pese a que el segundo *dataset* tiene casi 100 datos menos que el primero, ambos tienen aproximadamente el mismo número de diagnósticos negativos. Para ser exactos, en el primer *dataset* los casos positivos suponen el 32,65 % de los datos, mientras que en el segundo suponen el 22,13 %. Teniendo en cuenta que la prevalencia del SOP es del 5 % - 20 %, el primer *dataset* tiene una mayor proporción de positivos que la población general, pero el segundo *dataset* es más similar a la prevalencia del SOP en la población general.

En la Figura 4.11 se presenta la distribución del tipo de ciclo en ambos *datasets*. En este caso se puede observar que ambos *datasets* presentan más casos con tipo de ciclo 'Regular' que 'Irregular', lo que coincide con el criterio de diagnóstico del conocimiento experto. También se observa como la proporción de resultados es similar en ambos *datasets*.

En la Figura 4.12 se presenta la distribución de la variable "Pérdida de cabello" en ambos *datasets*. En este caso se puede observar como la proporción de las variables es muy distinta entre los *datasets*: el primero tiene más casos negativos que positivos, mientras que en el segundo es justamente al contrario.

En la Figura 4.13 se representa la distribución de la variable "Crecimiento de pelo" de ambos *datasets*. En este caso se puede observar como ambos *datasets* contienen más casos negativos que positivos, y la proporción entre casos negativos y positivos resulta aparentemente similar en ambos casos.

Así mismo, para profundizar el análisis de las diferencias entre ambos *datasets*, se ha realizado

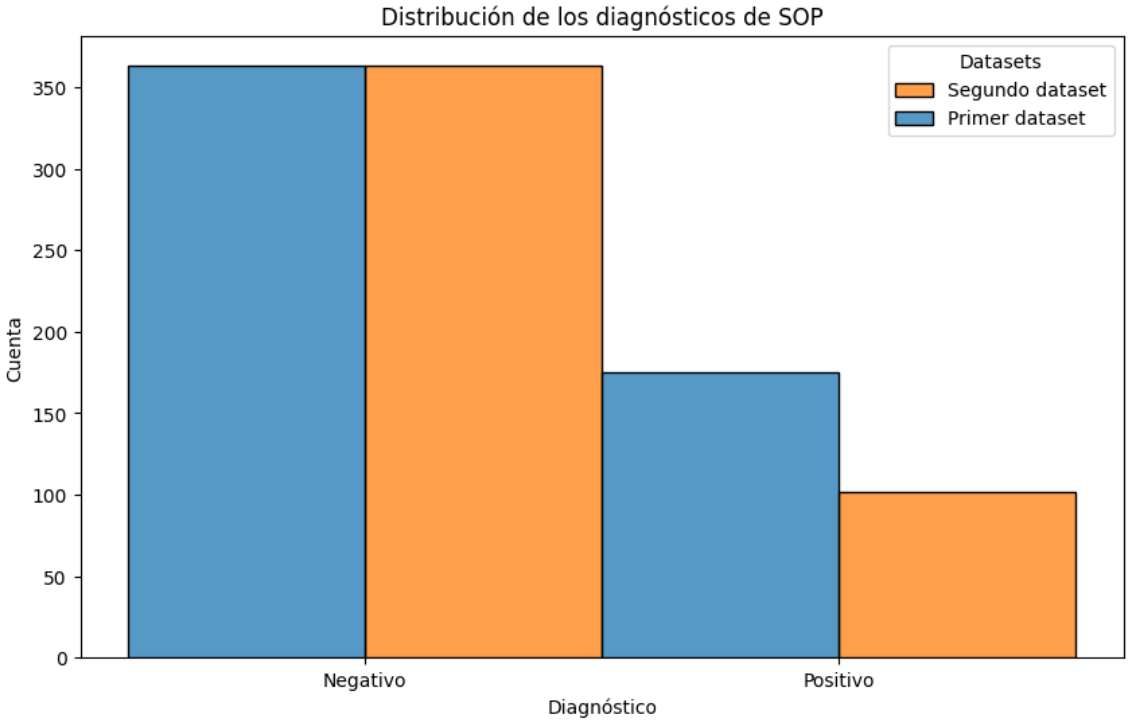

Figura 4.10: Número de diagnósticos negativos y positivos del primer (azul) y segundo (naranja) *dataset*.

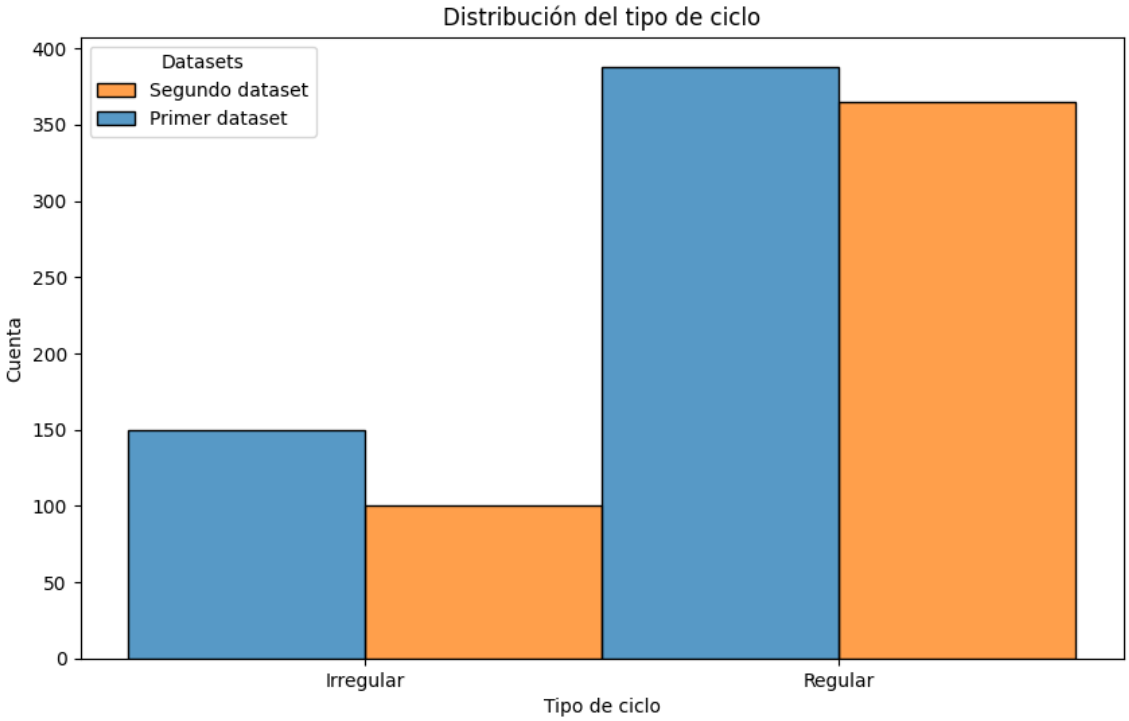

Figura 4.11: Distribución del tipo de ciclo del primer (azul) y segundo (naranja) *dataset*.

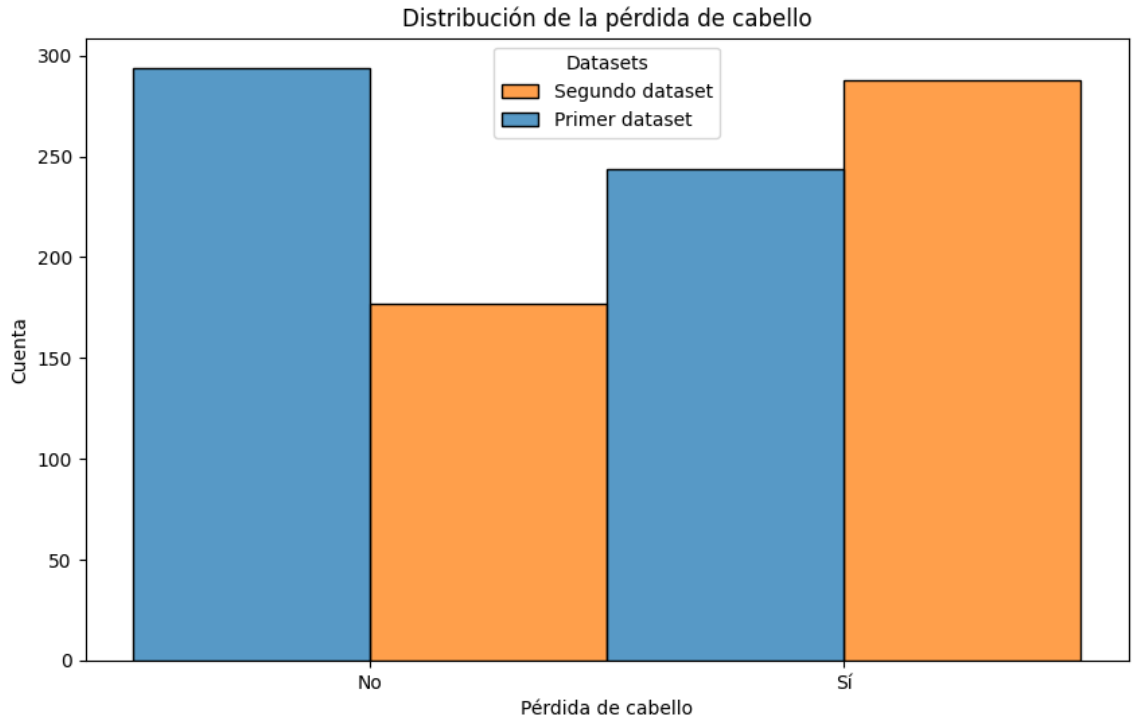

Figura 4.12: Distribución de la pérdida de cabello en el primer (azul) y segundo (naranja) *dataset*.

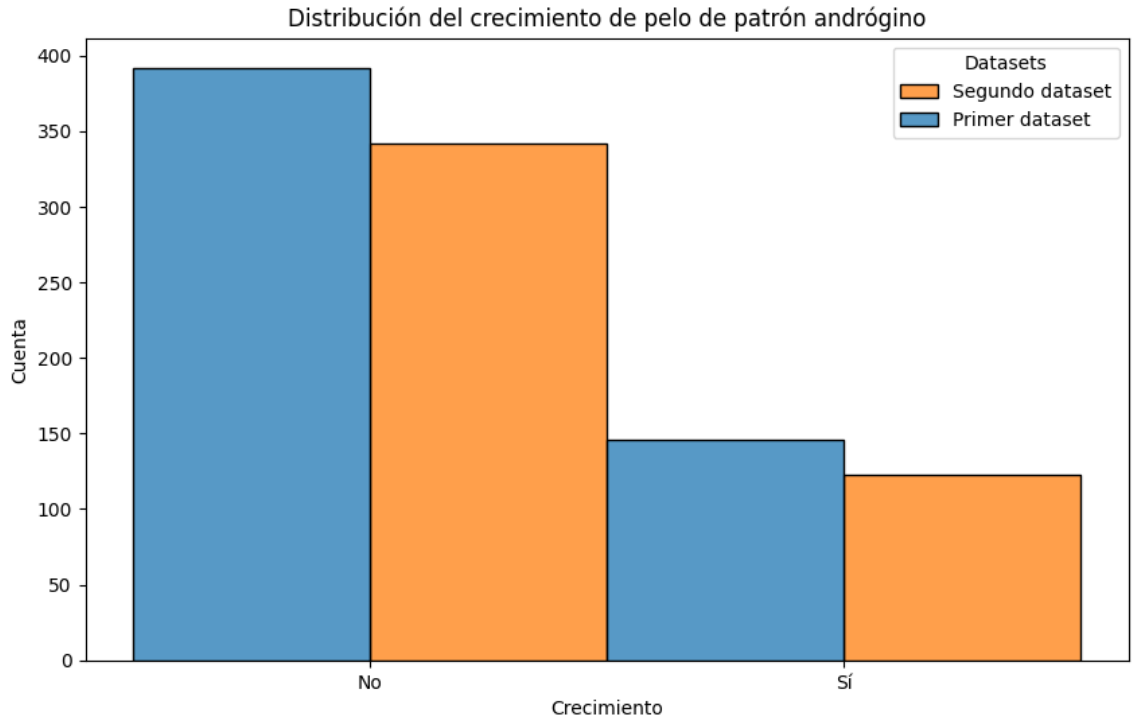

Figura 4.13: Distribución del crecimiento de pelo de patrón andrógino en el primer (azul) y segundo (naranja) *dataset*.

un test de Kolmogórov-Smirnov, comparando la distribución de ambos *datasets* en cada variable. De este forma se obtuvieron los resultados de la Tabla 4.4.

Feature	**KS-statistic**	**p-value**
Weight	0.06	0.40
Height	0.24	5.76e-13
Blood Group	0.17	8.51e-7
Weight gain	0.12	1.51e-3
Hair growth	4.56e-3	1
Skin darkening	0.04	0.89
Hair loss	0.17	1.12e-6
Pimples	0.08	0.10
Fast food	0.16	8.10e-6
Reg. Exercise	0.03	0.92
PCOS	0.10	8.37e-3
Cycle(R/I)	0.06	0.28
BMI	0.16	7.96e-6

Tabla 4.4: Resultados del test de Kolmogórov-Smirnov de ambos *datasets*.

De esta forma, podemos ver que hay variables que siguen distribuciones similares en ambos *datasets*, como el crecimiento de pelo de patrón andrógino o el ejercicio regular; y otras que son totalmente distinguibles, como pueden ser la altura o el grupo sanguíneo.

Para profundizar esta comparación entre ambos *datasets* se ha realizado una validación de adversarios o *Adversarial Validation (AV)*. Esto consiste en juntar los datos en un único *dataset*, dando una etiqueta a los datos procedentes de cada *dataset* (0 para el primero y 1 para el segundo), y crear un clasificador con un algoritmo (en este caso se utilizó *XGBClassifier*), de forma que sin utilizar la etiqueta que marca cada *dataset*, si ambos contienen la misma información deberían ser indistinguibles, y por tanto el clasificador no lo podrá clasificar. De esta forma se puede realizar un test, como por ejemplo el área de la curva ROC, y si ambos *datasets* son totalmente indistinguibles se obtendrá un resultado de en torno a 0,5. Si, al contrario, los *datasets* son distintos y el clasificador los puede diferenciar fácilmente, el área de la curva ROC será 1.

De esta forma, utilizando las variables comunes entre ambos *datasets*, se obtuvo un área de la curva ROC de 0,743047, lo que indica que ambos *datasets* son distinguibles. Cabe destacar que para realizar esta prueba se decidió eliminar las variables que se conocía con anterioridad que eran diferentes, como son la edad, la altura, el peso y el índice de masa corporal. De esta forma, las variables que se utilizaron para esta prueba son: grupo sanguíneo, aumento de peso, crecimiento de pelo de patrón andrógino, oscurecimiento de piel, caída del cabello, acné, comida rápida, ejercicio regular, diagnóstico de SOP y tipo de ciclo. Debido a esto resultado, se observó la importancia de las variables que se habían utilizado en el clasificador, obteniendo el siguiente resultado:

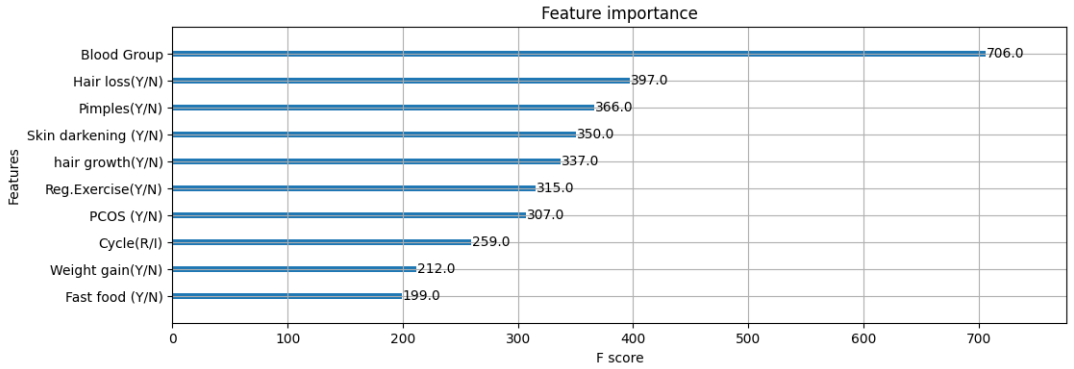

Figura 4.14: Importancia de las variables utilizadas en el clasificador de la validación de adversarios.

Como se puede observar, la variable con mayor importancia en el clasificador es el grupo sanguíneo. A continuación se puede observar un histograma de la variable grupo sanguíneo en ambos *datasets*:

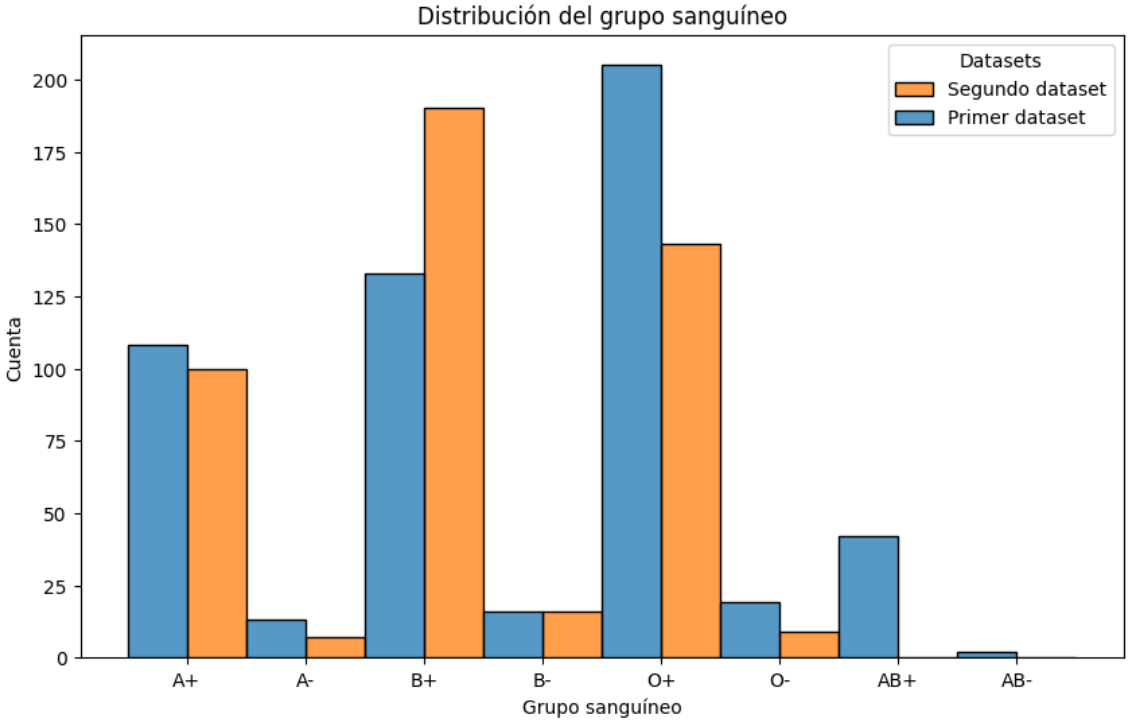

Figura 4.15: Distribución del grupo sanguíneo en el primer (azul) y segundo (naranja) *dataset*.

Como se puede observar, el grupo sanguíneo más común en el primer *dataset* es el O+, mientras que en el segundo *dataset* es el B+. Esto podría indicar que los grupos de pacientes son distintos en cuanto al origen étnico, ya que el grupo sanguíneo mayoritario es distinto para etnias distintas [40].

A continuación se decidió repetir el test de validación de adversarios, analizando únicamente las variables de ambos *datasets* que se utilizan para diagnosticar el Síndrome de Ovario Poliquístico, para tratar de determinar si hay alguna diferencia entre ambos *datasets* en cuanto a la forma de diagnosticar el SOP. Para ello se utilizaron solamente las variables: crecimiento de pelo de patrón andrógino, caída del cabello, oscurecimiento de piel, acné, tipo de ciclo y aumento de peso. De esta forma se obtiene un área de la curva ROC de 0,640976. Por tanto, se puede

determinar que ambos *datasets* son ligeramente distintos en cuanto a las variables de diagnóstico del Síndrome de Ovario Poliquístico.

4.4. Conclusiones

Se ha analizado un primer *dataset* que contiene información de entorno a 500 pacientes. Este *dataset* contiene tanto información básica de síntomas de las pacientes, como resultados de pruebas médicas como ecografías ováricas y analíticas sanguíneas. Este *dataset* tiene una proporción de diagnósticos positivos superior a la proporción de la población general, y la correlación entre sus variables predictoras y la variable objetivo concuerda con el conocimiento médico sobre el diagnóstico del SOP. Así mismo, se ha analizado un segundo *dataset* que contiene información de entorno a 400 pacientes, pero que no contenía resultados de pruebas médicas. Este segundo *dataset* tenía una proporción de positivos ligeramente superior a la población general, y la correlación de sus variables predictoras con el diagnóstico de SOP también coincidía con el conocimiento médico experto. Se han realizado varias pruebas para comparar ambos *datasets*, y se ha observado que tienen diferencias significativas respecto a varias variables, como la edad y el grupo sanguíneo. Sin embargo, se ha comprobado que en cuanto a las variables que se utilizan para diagnosticar el Síndrome de Ovario Poliquístico, la distribución de las variables es muy similar, lo que parece indicar que aunque la información provenga de pacientes distintas, el criterio para diagnosticar el Síndrome de Ovario Poliquístico es similar.

De esta forma, se ha determinado que ambos *datasets* pueden resultar útiles para desarrollar el modelo de diagnóstico que se ha propuesto. En particular, el primer *dataset* se puede utilizar para desarrollar los tres modelos de diagnóstico que se habían considerado: uno con datos básicos y síntomas de la paciente, un segundo modelo con resultados de pruebas ecográficas, y un tercer modelo con resultados de analíticas sanguíneas. Igualmente, se ha determinado que el segundo *dataset* puede resultar útil para desarrollar un cuarto modelo, similar al primer modelo que utiliza solamente síntomas y datos básicos de la paciente.

Capítulo 5

Desarrollo del modelo

5.1. Preprocesado

Una vez determinados los datos que se iban a utilizar, se realizó un proceso de preprocesado de los mismos para tratar de obtener la mayor información posible de estos *datasets*. Para ello, en primer lugar se observó que en el primer *dataset* la variable correspondiente al índice de masa corporal estaba en blanco en muchas de las filas, por lo que se decidió utilizar las variables de altura y peso para rellenar esos valores faltantes. Así mismo, en el segundo *dataset* no existía esta variable de índice de masa corporal, pero al tener la información correspondiente a la altura y el peso, se pudo crear esta variable.

Igualmente, en el primer *dataset* había dos variables que tenían muchas filas en blanco: las variables correspondientes a la proporción cadera/cintura y la proporción de hormonas FSH/LH. Al tener la información de esas cuatro variables en todas las filas, se pudo rellenar esta información fácilmente. Así mismo, este *dataset* contenía tres filas con algún campo con valores no numéricos (NaN), que se eliminaron del conjunto de datos para poder proceder con el desarrollo del modelo. También se transformaron algunas variables que contenían números enteros a números flotantes para facilitar el desarrollo del modelo, y se transformó la variable categórica de tipo de ciclo, que contenía la opción de ciclo 'Regular' representada con el número 2, y la opción de ciclo 'Irregular' con el número 4. Estas variables se transformaron a los valores 1 y 0 respectivamente, por similitud con el segundo *dataset*.

Una vez realizada esta parte del preprocesado, se dividió el primer *dataset* en tres conjuntos distintos, para realizar los tres modelos de diagnóstico que se han considerado. De esta forma, se creo un *dataset* que contenía solamente información básica de la paciente y su sintomatología; un segundo *dataset* con la información básica y los resultados de pruebas ecográficas; y un tercer *dataset* con toda la información original, es decir esa información y además los resultados de analíticas sanguíneas.

5.2. Reducción de variables

Después de realizar las primeras pruebas con distintos algoritmos, se observó que en algunos casos el tiempo de computación era excesivo, lo que podía dificultar el desarrollar una aplicación que de un diagnóstico certero a la paciente en tiempo real. Debido a esto se decidió realizar una reducción de variables, teniendo en cuenta por una parte el conocimiento médico del diagnóstico del Síndrome de Ovario Poliquístico, y por otra la importancia de las variables en el clasificador que mejores resultados obtuviese en ese momento. De esta forma, se redujo el número de variables del primer *dataset* de la siguiente manera: en el primer conjunto de datos, se redujo de 24 a 14 variables; en el segundo no se redujeron las variables de las pruebas ecográficas, lo que resultó en un total de 19 variables; y en el tercero se redujo de 41 a 24 variables. En cuanto al segundo *dataset*, no se realizó ninguna reducción de variables, ya que solo contenía 16 variables predictoras

DB (1)	DB (2)	DB + eco	DB + eco + analítica
Edad	Edad	Edad	Edad
Peso	Peso	Peso	Peso
Altura	Altura	Altura	Altura
BMI	BMI	BMI	BMI
Tipo de ciclo	Tipo de ciclo	Tipo de ciclo	Tipo de ciclo
Cadera		Cadera	Cadera
Cintura		Cintura	Cintura
Cadera/Cintura		Cadera/Cintura	Cadera/Cintura
Aumento de peso	Aumento de peso	Aumento de peso	Aumento de peso
Crecimiento de pelo	Crecimiento de pelo	Crecimiento de pelo	Crecimiento de pelo
Oscurecimiento de piel	Oscurecimiento de piel	Oscurecimiento de piel	Oscurecimiento de piel
Caída de pelo	Caída de pelo	Caída de pelo	Caída de pelo
Acné	Acné	Acné	Acné
Comida rápida	Comida rápida	Comida rápida	Comida rápida
	Grupo sanguíneo		
	Meses periodo		
	Ejercicio		
	Cambios de humor		
	Duración del ciclo		
		N. Fol. (I)	N. Fol. (I)
		N. Fol. (D)	N. Fol. (D)
		T. medio Fol. (I)	T. medio Fol. (I)
		T. medio Fol. (D)	T. medio Fol. (D)
		Endometrio	Endometrio
			LH
			FSH
			FSH/LH
			TSH
			AMH

Tabla 5.1: Variables de cada uno de los cuatro *datasets*.

y el tiempo de computación no era tan excesivo. En la Tabla 5.1 se puede observar las variables predictoras que se utilizaron en cada conjunto de datos.

5.3. Integración del conocimiento médico

Para desarrollar este modelo de diagnóstico se utilizó un sistema híbrido, combinando el modelo de Aprendizaje Automático con el conocimiento médico. Esto se hace desarrollando un sistema de reglas que, en este caso, separe los casos que pueden ser diagnosticados directamente con el conocimiento médico del Síndrome de Ovario Poliquístico; de los casos indecisos, que serán diagnosticados utilizando un algoritmo de Aprendizaje Automático. De esta forma se desarrollaron dos reglas distintas: una para el modelo que utiliza solamente los datos básicos y síntomas de la paciente; y una para el modelo que incluye las pruebas de ecografía y de analítica. Estas reglas se realizaron teniendo en cuenta el criterio médico de diagnóstico del SOP, y su traducción en las variables de ambos *datasets*.

En la Figura 5.1 se puede observar un esquema que representa el funcionamiento de la primera de estas reglas. Estas reglas representan cómo se transforman las reglas de diagnóstico médico en las variables del *dataset* que se ha utilizado. En este caso este sistema de reglas se aplica a los dos modelos de diagnóstico con datos básicos de las pacientes. En la Figura 5.2 se

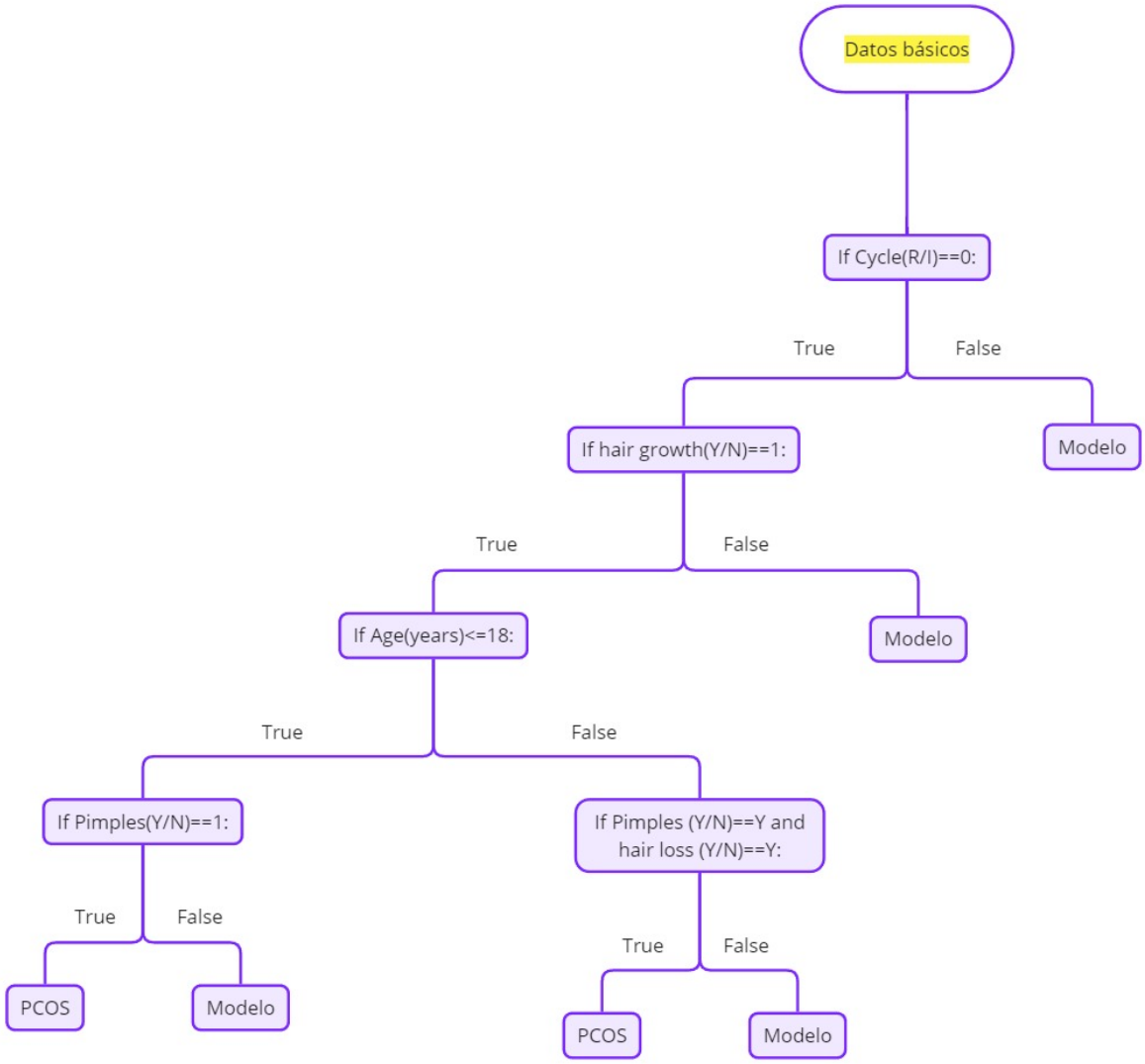

Figura 5.1: Esquema del sistema de reglas del primer modelo de diagnóstico.

puede observar el esquema correspondiente a los otros dos modelos. En estos dos casos se aplica el mismo sistema de reglas, ya que el modelo con resultados de analíticas no añade información que pueda proporcionar un diagnóstico directo.

5.4. Validación

Una parte fundamental del desarrollo de modelos de Aprendizaje Automático es la validación de los resultados. Esto se hace para reducir el sobre-ajuste o *overfitting*, que se produce cuando un modelo se entrena directamente con un conjunto de datos, de forma que el modelo se ajusta mucho a ese conjunto, pero a la hora de evaluarlo con datos distintos el resultado no es tan bueno como se esperaba, ya que ese modelo se ha 'sobre-ajustado' a los datos de entrenamiento. Esto se evita dividiendo el conjunto de datos que se va a utilizar en dos partes: una para entrenar el modelo, que se denomina conjunto de *train*, y otra para evaluar la calidad del modelo, que se denomina conjunto de *test*. De esta forma, cuando se utilizan métricas para evaluar la calidad de un modelo, estas métricas se aplican al conjunto de *test*, lo que nos da una mejor idea de cómo se va a comportar el modelo cuando se utilice para clasificar nuevos datos.

En este caso, para dividir los conjuntos de datos se utilizó la técnica *KFold*, con $k = 10$.

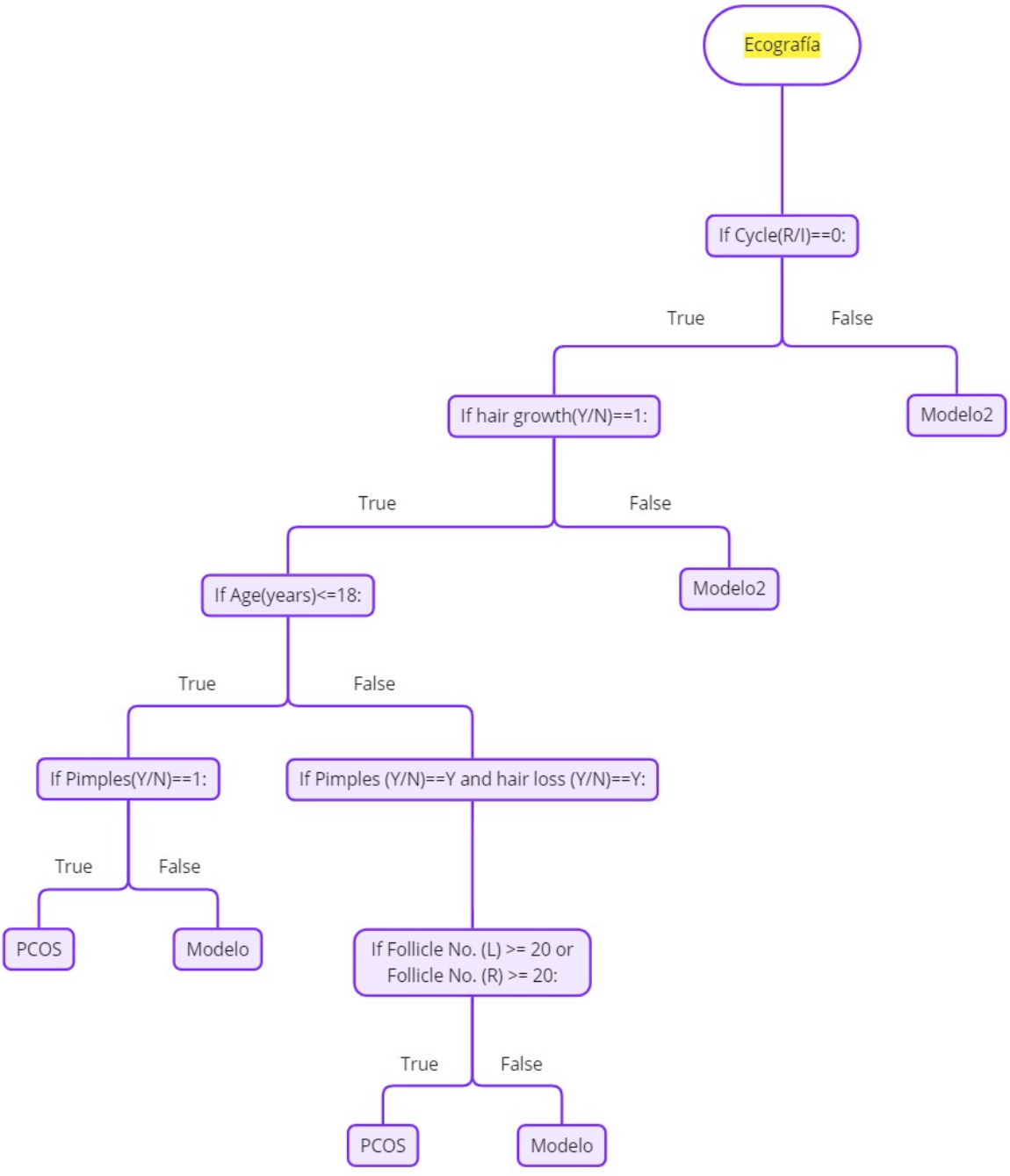

Figura 5.2: Esquema del sistema de reglas del segundo y tercer modelo de diagnóstico.

Esta técnica consiste en dividir los datos en 10 conjuntos de igual tamaño (1/10 del tamaño total de los datos), y entrenar el modelo correspondiente de Aprendizaje Automático un total de 10 veces, cada una utilizando uno de esos subconjuntos como el conjunto de *test*, y los otros 9 como conjunto de *train*. Se decidió utilizar esta técnica ya que en el caso de conjuntos de datos pequeños como el que hemos utilizado, suele ser más adecuada.

5.5. Proceso de entrenamiento

Una vez divididos los *datasets*, se entrenaron los modelos de Aprendizaje Automático escogidos, utilizando las funciones correspondientes en *Python*. En el caso del algoritmo *Decision tree*, se utilizó como criterio el índice Gini, y no se puso ninguna restricción a la profundidad del árbol. En el caso del modelo *Random Forest*, se determinó el número máximo de árboles a desarrollar en 100, como criterio se utilizó el índice Gini y no se determinó ninguna otra restricción.

En el caso del algoritmo *Support Vector Machine*, se utilizó un *Suppor Vector Classifier* con *kernel* lineal. Así mismo, se utilizó la técnica de *Grid Search* o búsqueda en cuadrícula para encontrar el valor óptimo del hiperparámetro C para este problema.

5.6. Resultados obtenidos

En la Tabla 5.2 se presentan los resultados de la métrica de *accuracy* o precisión de cada modelo. Esta métrica determina el porcentaje de casos en los que la predicción del modelo coincide con la variable objetivo.

Modelo	**DB (1)**	**DB (2)**	**DB + eco**	**DB + eco + analítica**
Decision Tree	96.7	97.6	98.0	97.6
Random Forest	96.7	97.8	99.1	99.1
SVM	85.3	84.4	92.2	92.0

Tabla 5.2: *Accuracy* de cada modelo utilizado. De izquierda a derecha, se presentan los modelos realizados con los datos básicos del primer *dataset*, los datos básicos del segundo *dataset*, los datos básicos y de ecografía del primer *dataset* y los datos básicos, de ecografía y de analítica del primer *dataset*.

Como se puede observar, los modelos tienen menor *accuracy* en el caso de los datos básicos, que es el conjunto con menor número de variables, y estos mejoran con el aumento del número de variables, en el caso de los datos básicos y la ecografía. Sin embargo, en el último caso de los datos básicos con ecografía y analítica, en algunos casos la *accuracy* disminuye. Esto se debe a que este conjunto incluye muchas variables que no se utilizan para diagnosticar el Síndrome de Ovario Poliquístico, y solo una hormona que sí se utiliza en algunas circunstancias, aunque generalmente se usan los datos de los dos conjuntos de datos anteriores. Debido a esto, el último conjunto de datos añade un gran número de variables predictoras que realmente no están fuertemente relacionadas con la variable objetivo, y eso hace que la predicción del modelo sea más errónea. Así mismo, se puede observar que el modelo que produce mejor *accuracy* en la mayoría de los casos es *Random Forest*. En el caso de *Decision tree*, produce mejores resultados que *SVM* en todos los modelos, que es el algoritmo que produce peores resultados.

Para ampliar el análisis de las predicciones de estos modelos, se ha analizado el valor F o *F score*, que se obtiene en función del número de verdaderos positivos (TP), falsos positivos (FP), y falsos negativos (FN), de esta manera:

$$F = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

En la Tabla 5.3 se presentan los resultados de la métrica F.

Modelo	DB (1)	DB (2)	DB + eco	DB + eco + analítica
Decision Tree	94.9	94.7	96.9	96.3
Random Forest	94.9	95.1	98.6	98.6
SVM	77.0	64.4	87.9	87.5

Tabla 5.3: *F score* de cada modelo utilizado. De izquierda a derecha, se presentan los modelos realizados con los datos básicos del primer *dataset*, los datos básicos del segundo *dataset*, los datos básicos y de ecografía del primer *dataset* y los datos básicos, de ecografía y de analítica del primer *dataset*.

Esta métrica nos da una idea de la proporción de verdaderos positivos y falsos positivos y negativos. Se puede observar una tendencia similar al caso de la *accuracy*, el modelo *Random Forest* es el que produce mejores resultados. Esta métrica también puede ayudar a comprobar si la proporción de clases en la clasificación es la adecuada, o el modelo se decanta más por una clase en particular. Al tener resultados en todos los casos de entre 94 % y 98 %, podemos comprobar que la proporción de las clases está balanceada. Sin embargo, en el caso del algoritmo *SVM*, se observa que hay un mayor número de falsos positivos y negativos, ya que en el mejor de los casos el valor F no llega al 90 %.

Así mismo, se ha obtenido también la curva ROC de cada modelo. En la Figura 5.3 se pueden observar las curvas ROC de cada modelo desarrollado con el algoritmo *Decision tree*. En este caso se puede observar como los cuatro modelos producen una clasificación muy buena, y se aprecia cierta mejoría en los dos últimos modelos frente a los dos primeros.

En la Figura 5.4 se pueden observar las curvas ROC de cada modelo desarrollado con el algoritmo *Random Forest*. En este caso se observan resultados muy similares al apartado anterior: los cuatro modelos producen buenos clasificadores, aunque la clasificación de los dos últimos modelos es mejor que la de los dos primeros.

En la Figura 5.5 se pueden observar las curvas ROC de cada modelo desarrollado con el algoritmo *Support Vector Machine*. En este caso se observa que la clasificación es significativamente peor que en los dos algoritmos anteriores. En particular se ha obtenido un valor de verdaderos positivos notablemente menor que en los otros algoritmos. También se observa una mejoría en los dos últimos modelos respecto a los dos primeros.

Esta métrica nos proporciona información sobre la precisión de clasificación en eventos binarios o no binarios. En este caso, proporciona unos resultados muy similares a los que se han obtenido con la *accuracy*: el área aumenta generalmente al aumentar el número de variables, salvo en el caso del modelo con los resultados de la analítica sanguínea. Así mismo, el modelo que genera los mejores resultados en la mayoría de los casos es *Random Forest*. De igual forma, el algoritmo *SVM* es el que produce peores resultados en la clasificación.

Con estas métricas se ha podido observar los resultados de la clasificación de cada modelo. Con la *accuracy* podemos observar la precisión general de cada modelo, mientras que con el valor F vemos la proporción de verdaderos positivos respecto a los falsos positivos y negativos. Igualmente, con la curva ROC podemos observar de forma gráfica la precisión del clasificador binario.

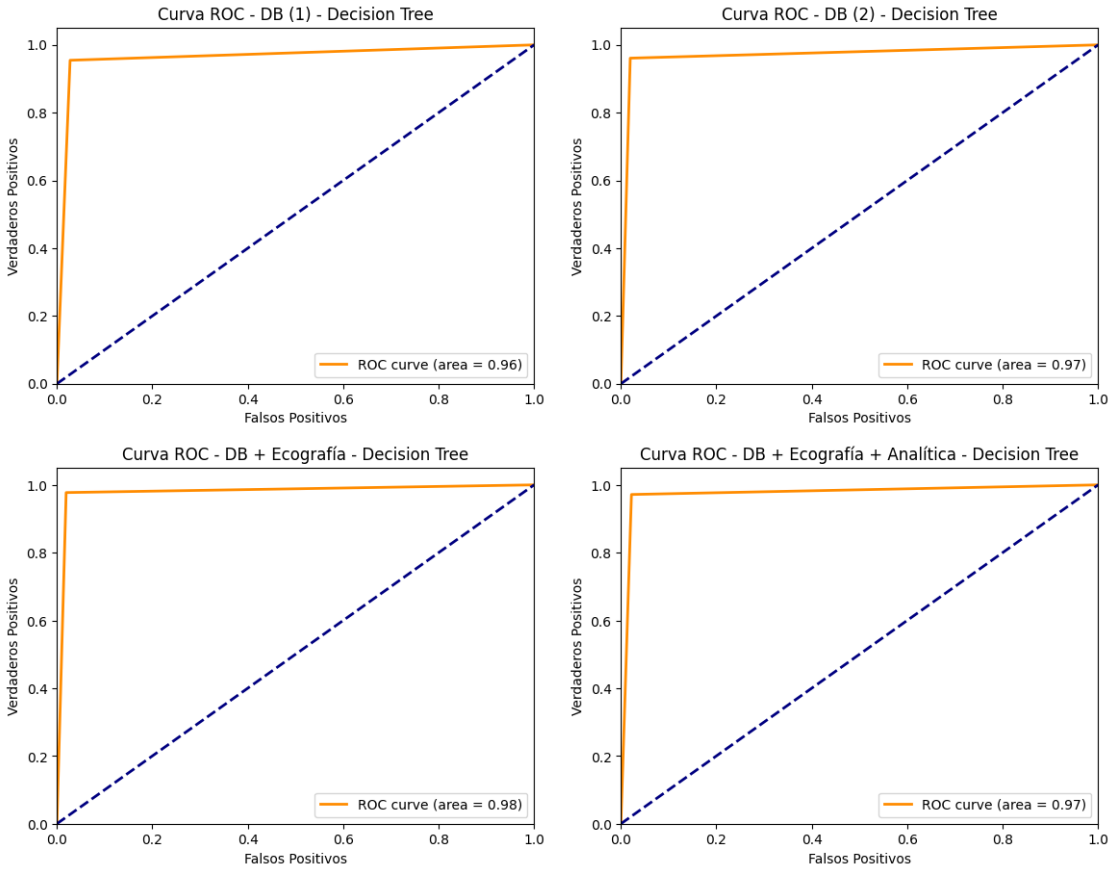

Figura 5.3: Curvas ROC de los cuatro modelos desarrollados con el algoritmo *Decision tree*.

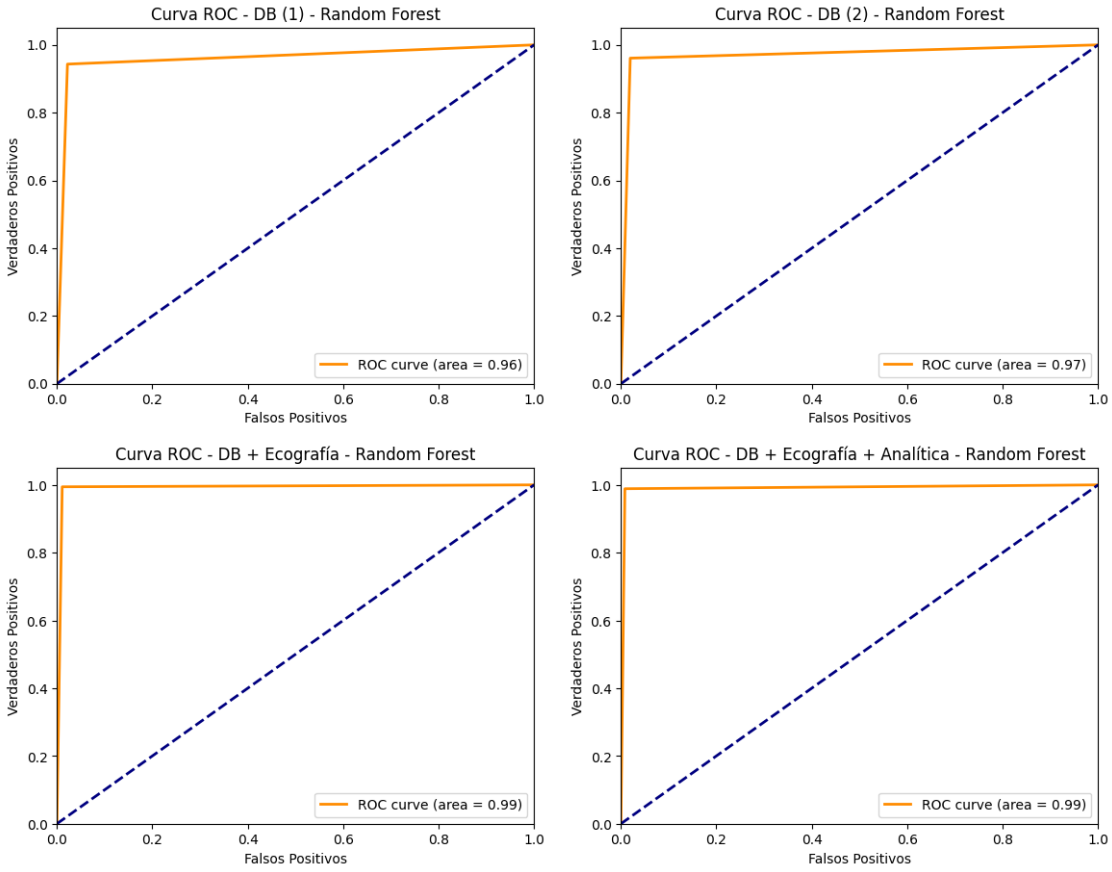

Figura 5.4: Curvas ROC de los cuatro modelos desarrollados con el algoritmo *Random Forest*.

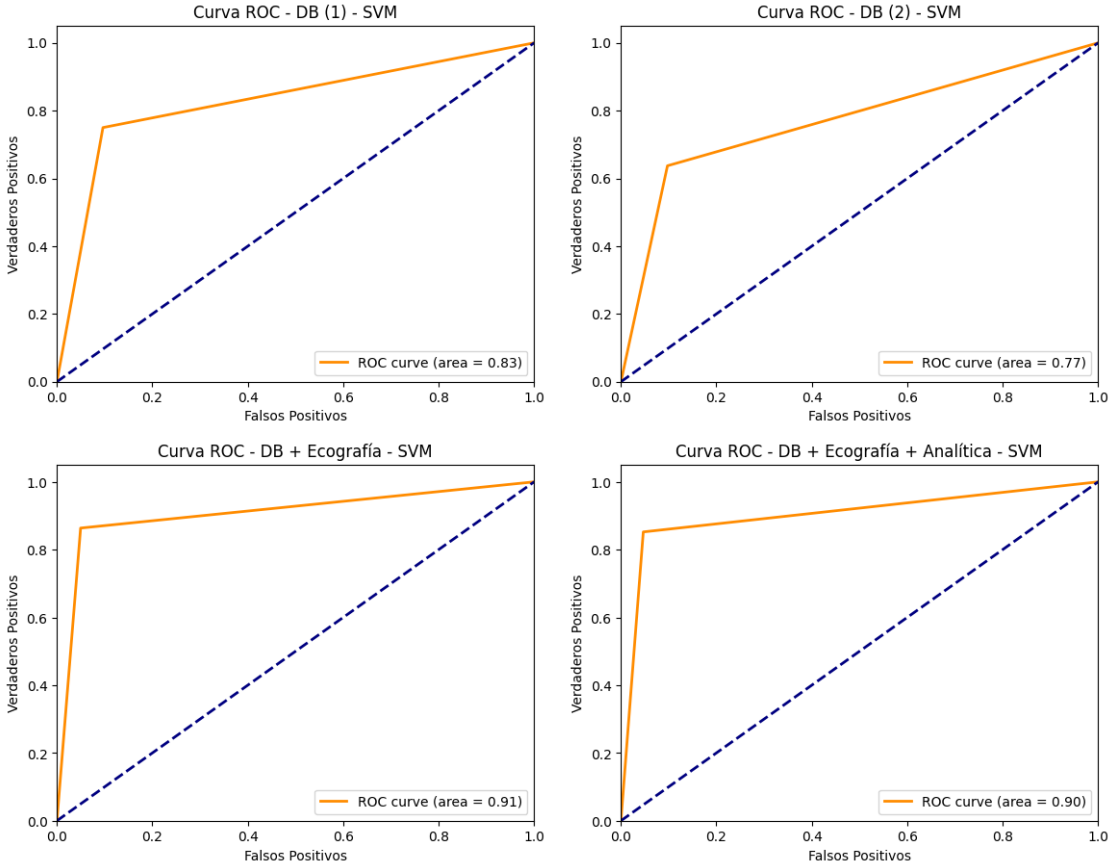

Figura 5.5: Curvas ROC de los cuatro modelos desarrollados con el algoritmo *Support Vector Machine*.

Capítulo 6

Desarrollo del Producto Mínimo Viable (MVP)

6.1. Diseño de Intellcyst

Para diseñar la aplicación se utilizó *Gradio*, una herramienta para desarrollar aplicaciones a partir de modelos de Aprendizaje Automático [41]. Puede ver una versión de la aplicación aquí: `https://intellcyst.deducedata.solutions/` (usuario: deduce, contraseña: DDS-makeitsmart).

En primer lugar se dividió la aplicación en los cuatro apartados principales que iba a utilizar el usuario: el modelo de diagnóstico con datos básicos, el modelo de diagnóstico con ecografía, el modelo de diagnóstico con analítica y el tratamiento personalizado. Así mismo, se generó un "número de paciente" a partir de un número aleatorio de 7 cifras, para poder guardar los resultados del modelo más adelante, identificando los datos de cada paciente pero respetando su privacidad.

En el apartado del primer diagnóstico, aparecen los campos correspondientes para introducir las variables de diagnóstico. En el caso de las variables numéricas aparece un campo donde escribir el número de la variable; mientras que en el caso de las variables categóricas aparecen dos botones, con los valores correspondientes a esa variable categórica. A continuación de estos campos, aparece un botón donde calcular el resultado del diagnóstico, y un botón para guardar los resultados y utilizarlos más adelante. El resultado del diagnóstico es tanto categórico, es decir si el diagnóstico del modelo es positivo o negativo, como la probabilidad de que este diagnóstico sea correcto. En el caso de que la probabilidad no supere un umbral (que se ha establecido en el 70 %), aparecerá un mensaje indicando que el diagnóstico no es lo bastante seguro, y se recomienda hacer una prueba de ecografía para comprobar el diagnóstico. En el caso de que la probabilidad de diagnóstico supere este umbral, aparecerá un mensaje que indica que el diagnóstico es seguro, y que puede saltar al apartado de tratamiento.

Así mismo, después del resultado aparece un apartado desplegable, donde se pueden introducir unas "variables opcionales", que son las variables que se retiraron del modelo en la reducción de variables. De esta forma si la paciente o el personal sanitario tiene esa información y quiere añadirla para comprobar si el diagnóstico es distinto, puede hacerlo. En la Figura 6.1 se puede observar una captura de pantalla de la aplicación.

En el caso de los apartados de ecografía y analítica, el funcionamiento es similar al apartado anterior: en primer lugar aparecen los campos para introducir las variables correspondientes, un botón para calcular el diagnóstico, un apartado de resultado donde aparece el diagnóstico, la probabilidad de diagnóstico y la recomendación pertinente; un botón para guardar el resultado obtenido y un apartado de "variables opcionales". En las Figuras 6.2 y 6.3 se pueden observar capturas de pantallas de estos apartados de la aplicación.

Finalmente, se encuentra el apartado de tratamiento. En este apartado primero hay que

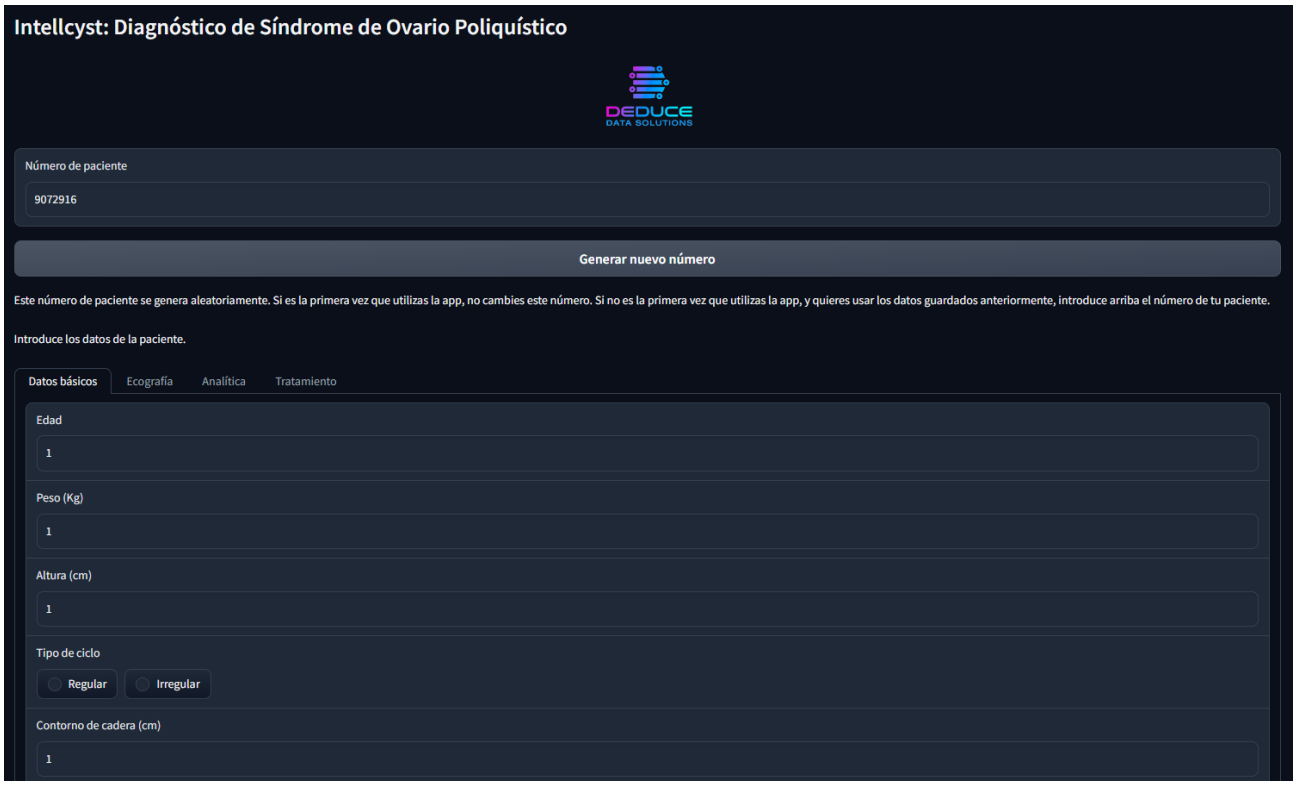

Figura 6.1: Captura de pantalla de la aplicación Intellcyst.

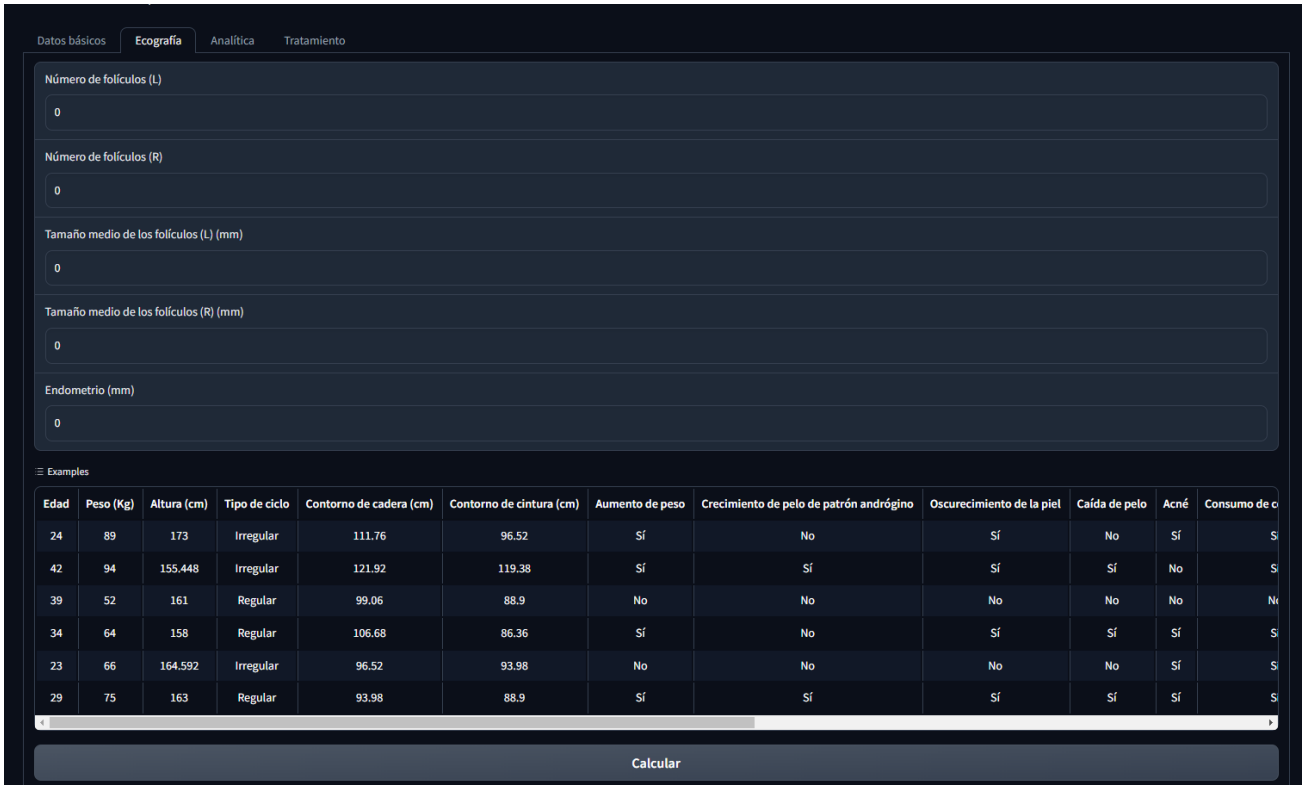

Figura 6.2: Captura de pantalla de la aplicación Intellcyst.

Figura 6.3: Captura de pantalla de la aplicación Intellcyst.

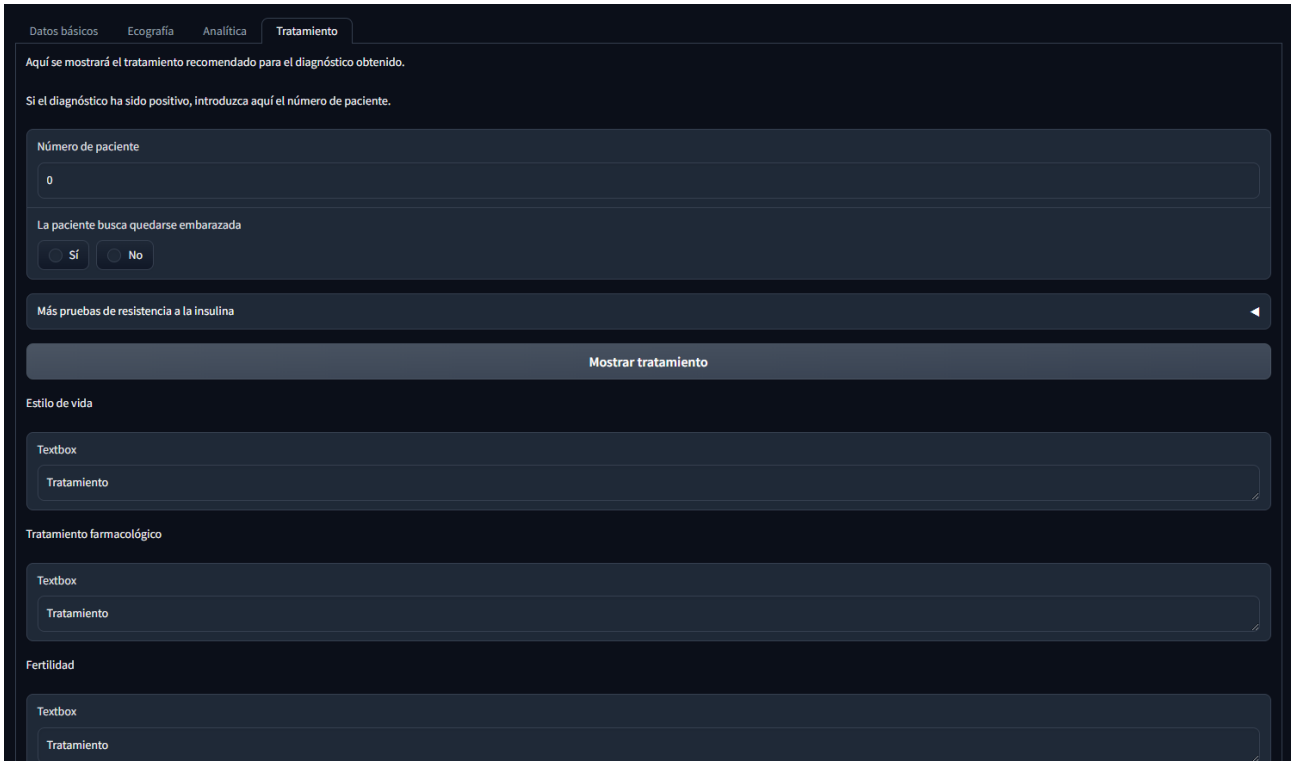

Figura 6.4: Captura de pantalla de la aplicación Intellcyst.

introducir el número de paciente que se ha utilizado para guardar los resultados, para que la aplicación utilice los datos de la paciente para determinar un tratamiento personalizado. También aparece un campo donde la paciente puede introducir si está en un proceso de búsqueda de embarazo, y otro campo donde se pueden incluir resultados de otras pruebas extras de resistencia a la insulina. Este apartado se ha incluido especialmente para aquellos casos en los que no se ha realizado la analítica sanguínea, y así, si el personal sanitario tiene esta información, la aplicación puede proporcionar un tratamiento más adecuado. Una vez establecidas estas variables, la aplicación desarrolla un tratamiento especializado, dividido en cuatro campos: estilo de vida, tratamiento farmacológico, tratamiento de fertilidad y tratamiento psicológico. Este tratamiento es totalmente específico para la información que la aplicación tiene de la paciente en concreto, y además incluye información sobre el Síndrome de Ovario Poliquístico que puede ser útil para la paciente. Así mismo, hay un botón desde el cual la paciente puede descargar una copia en formato *pdf* de este tratamiento. En la Figura 6.4 se puede observar una captura de pantalla de este apartado de la aplicación.

6.2. Integración del modelo

Debido a los resultados de los modelos estudiados, se decidió utilizar el algoritmo *Random Forest* en el desarrollo de la aplicación. De esta forma, una vez entrenados los modelos con los datos correspondientes, la aplicación utiliza los datos que se han introducido para obtener un diagnóstico, junto con la probabilidad de que este diagnóstico sea correcto. Para ello utiliza un sistema híbrido que combina el conocimiento experto junto con el modelo de Aprendizaje Automático correspondiente.

Capítulo 7

Conclusiones

Este trabajo se ha centrado en diseñar una herramienta que proporcione un acompañamiento a las pacientes con Síndrome de Ovario Poliquístico, y en facilitar el trabajo de los profesionales sanitarios que se enfrentan al problema de diagnosticar y tratar esta enfermedad.

En primer lugar, para realizar este trabajo se ha tenido que realizar un estudio exhaustivo sobre el Síndrome de Ovario Poliquístico, así como su diagnóstico y su sintomatología. Se ha estudiado la literatura existente en cuanto al diagnóstico, así como los estudios sobre el uso de Aprendizaje Automático en el diagnóstico médico.

Así mismo, se han presentado dos *datasets* que contienen información relevante sobre pacientes con y sin Síndrome de Ovario Poliquístico; y se han analizado sus características y las diferencias entre ambos. Se ha comprobado que los dos *datasets* son distintos, pero que ambos pueden resultar útiles para desarrollar un modelo de diagnóstico del SOP.

Igualmente se han desarrollado una serie de modelos de Aprendizaje Automático, combinados con el conocimiento experto sobre el SOP, para tratar de determinar cuál producía mejores resultados en el diagnóstico. Se han desarrollado una serie de modelos de clasificación de Aprendizaje Automático, combinando también el conocimiento experto mediante un sistema híbrido. Se ha encontrado que el mejor modelo de diagnóstico se producía utilizando el algoritmo *Random forest*, obteniendo una *accuracy* del 99,1 % y un valor F del 98,6 %. Esto implica el modelo de diagnóstico funciona correctamente en la gran mayoría de los casos, y en los que no funciona correctamente, no se producen más falsos positivos que negativos o viceversa. Esto proporciona una gran confianza al modelo a la hora de utilizarlo como sistema de diagnóstico clínico.

De igual manera, se ha observado que el modelo realizado tiene ciertas limitaciones. Los datos que se han utilizado para entrenar los modelos solo contienen datos básicos y síntomas de pacientes menores de edad, por lo que en el caso del modelo con resultados de pruebas médicas para menores de edad, no podemos determinar completamente la eficacia del modelo. Así mismo, en el caso del modelo con datos básicos, la *accuracy* es ligeramente menor que en el resto de casos, en torno al 97 %. Igualmente, el número de pacientes que se han utilizado para entrenar cada modelo es de en torno a 500 pacientes.

De esta forma, para mejorar el modelo de diagnóstico sería adecuado utilizar más pacientes, y también algún *dataset* que contenga resultados de pruebas médicas de pacientes menores de edad.

Igualmente, se ha desarrollado una aplicación que permite a cualquier usuario utilizar el modelo de diagnóstico diseñado. Utilizando la herramienta *Gradio* se ha diseñado una aplicación sencilla de utilizar que permite al usuario obtener un diagnóstico en los tres casos distintos que se han estudiado; así como la probabilidad de que este diagnóstico sea correcto, y una serie de recomendaciones.

Así mismo, se ha utilizado el conocimiento experto sobre el Síndrome de Ovario Poliquístico para proporcionar un curso de tratamiento especializado y adecuado para cada paciente que utilice la aplicación.

Bibliografía

[1] Teede, H. J., Tay, C. T., Laven, J. J., Dokras, A., Moran, L. J., Piltonen, T. T., Joham, A. E. (2023). *Recommendations from the 2023 international evidence-based guideline for the assessment and management of polycystic ovary syndrome.* European journal of endocrinology, 189(2), G43-G64. https://doi.org/10.1210/clinem/dgad463

[2] Escobar-Morreale, H. F. (2018). *Polycystic ovary syndrome: definition, aetiology, diagnosis and treatment.* Nature Reviews Endocrinology, 14(5), 270-284. doi:10.1038/nrendo.2018.24

[3] Reproducción asistida ORG. *Síndrome del ovario poliquístico (SOP): causas, síntomas y tratamiento.* https://www.reproduccionasistida.org/sindrome-de-ovarios-poliquisticos/ [14-06-24]

[4] Stephen Franks. *Diagnosis of Polycystic Ovarian Syndrome: In Defense of the Rotterdam Criteria.* The Journal of Clinical Endocrinology & Metabolism, Volume 91, Issue 3, 1 March 2006, Pages 786–789, https://doi.org/10.1210/jc.2005-2501

[5] Sadeghi, H. M., Adeli, I., Calina, D., Docea, A. O., Mousavi, T., Daniali, M., Nikfar, S., Tsatsakis, A., & Abdollahi, M. (2022). *Polycystic Ovary Syndrome: A Comprehensive Review of Pathogenesis, Management, and Drug Repurposing.* International journal of molecular sciences, 23(2), 583. https://doi.org/10.3390/ijms23020583

[6] Islam, H., Masud, J., Islam, Y. N., & Haque, F. K. M. (2022). *An update on polycystic ovary syndrome: A review of the current state of knowledge in diagnosis, genetic etiology, and emerging treatment options.* Women's health (London, England), 18, 17455057221117966. https://doi.org/10.1177/17455057221117966

[7] Stener-Victorin, E., Padmanabhan, V., Walters, K. A., Campbell, R. E., Benrick, A., Giacobini, P., ... & Abbott, D. H. (2020). *Animal models to understand the etiology and pathophysiology of polycystic ovary syndrome.* Endocrine reviews, 41(4), bnaa010. https://doi.org/10.1210/endrev/bnaa010

[8] Azziz, R., Carmina, E., Chen, Z., Dunaif, A., Laven, J. S., Legro, R. S., ... & Yildiz, B. O. (2016). *Polycystic ovary syndrome.* Nature reviews Disease primers, 2(1), 1-18.

[9] Sanchez-Garrido, M. A., & Tena-Sempere, M. (2020). *Metabolic dysfunction in polycystic ovary syndrome: Pathogenic role of androgen excess and potential therapeutic strategies.* Molecular metabolism, 35, 100937. https://doi.org/10.1016/j.molmet.2020.01.001

[10] Lim, S. S., Davies, M. J., Norman, R. J., & Moran, L. J. (2012). *Overweight, obesity and central obesity in women with polycystic ovary syndrome: a systematic review and meta-analysis.* Human reproduction update, 18(6), 618–637. https://doi.org/10.1093/humupd/dms030

[11] Joham, A. E., Teede, H. J., Ranasinha, S., Zoungas, S., & Boyle, J. (2015). *Prevalence of infertility and use of fertility treatment in women with polycystic ovary syndrome: data from a large community-based cohort study.* Journal of women's health (2002), 24(4), 299–307. https://doi.org/10.1089/jwh.2014.5000

[12] Teede, H. J., Misso, M. L., Deeks, A. A., Moran, L. J., Stuckey, B. G., Wong, J. L., Norman, R. J., Costello, M. F., & Guideline Development Groups (2011). *Assessment and management of polycystic ovary syndrome: summary of an evidence-based guideline.* The Medical journal of Australia, 195(6), S65–S112. https://doi.org/10.5694/mja11.10915

[13] Azziz, R., Carmina, E., Dewailly, D., Diamanti-Kandarakis, E., Escobar-Morreale, H. F., Futterweit, W., Janssen, O. E., Legro, R. S., Norman, R. J., Taylor, A. E., Witchel, S. F., & Task Force on the Phenotype of the Polycystic Ovary Syndrome of The Androgen Excess and PCOS Society (2009). *The Androgen Excess and PCOS Society criteria for the polycystic ovary syndrome: the complete task force report.* Fertility and sterility, 91(2), 456–488. https://doi.org/10.1016/j.fertnstert.2008.06.035

[14] Roos, N., Kieler, H., Sahlin, L., Ekman-Ordeberg, G., Falconer, H., & Stephansson, O. (2011). *Risk of adverse pregnancy outcomes in women with polycystic ovary syndrome: population based cohort study.* BMJ (Clinical research ed.), 343, d6309. https://doi.org/10.1136/bmj.d6309

[15] Guan, C., Zahid, S., Minhas, A. S., Ouyang, P., Vaught, A., Baker, V. L., & Michos, E. D. (2022). *Polycystic ovary syndrome: a risk-enhancing"factor for cardiovascular disease.* Fertility and sterility, 117(5), 924–935. https://doi.org/10.1016/j.fertnstert.2022.03.009

[16] Dumesic, D. A., & Lobo, R. A. (2013). *Cancer risk and PCOS.* Steroids, 78(8), 782–785. https://doi.org/10.1016/j.steroids.2013.04.004

[17] Kevin P. Muprhy *Machine Learning. A probabilistic Perspective.* The MIT Press, 2012. ISBN: 978-0-262-01802-9

[18] Kingsford, C., & Salzberg, S. L. (2008). *What are decision trees?.* Nature biotechnology, 26(9), 1011–1013. https://doi.org/10.1038/nbt0908-1011

[19] Wang, Q. R., & Suen, C. Y. (1984). *Analysis and design of a decision tree based on entropy reduction and its application to large character set recognition.* IEEE Transactions on Pattern Analysis and Machine Intelligence, (4), 406-417.

[20] Rigatti, S. J. (2017). *Random forest.* Journal of Insurance Medicine, 47(1), 31-39. https://doi.org/10.17849/insm-47-01-31-39.1

[21] Ali, Z. A., Abduljabbar, Z. H., Taher, H. A., Sallow, A. B., & Almufti, S. M. (2023). *Exploring the power of eXtreme gradient boosting algorithm in machine learning: A review.* Academic Journal of Nawroz University, 12(2), 320-334.

[22] Margineantu, D. D., & Dietterich, T. G. (1997, July). *Pruning adaptive boosting.* In ICML (Vol. 97, pp. 211-218).

[23] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). *A comprehensive survey on support vector machine classification: Applications, challenges and trends.* Neurocomputing, 408, 189-215. https://doi.org/10.1016/j.neucom.2019.10.118

[24] Wu, Y. C., & Feng, J. W. (2018). *Development and application of artificial neural network.* Wireless Personal Communications, 102, 1645-1656. https://doi.org/10.1007/s11277-017-5224-x

[25] Pavlyshenko, B. (2018, August). *Using stacking approaches for machine learning models.* In 2018 IEEE second international conference on data stream mining & processing (DSMP) (pp. 255-258). IEEE.

[26] Elmannai, H., El-Rashidy, N., Mashal, I., Alohali, M. A., Farag, S., El-Sappagh, S., & Saleh, H. (2023).*Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and Explainable Artificial Intelligence.* Diagnostics (Basel, Switzerland), 13(8), 1506. https://doi.org/10.3390/diagnostics13081506

[27] Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., & Kumar, P. (2021). *Artificial intelligence to deep learning: machine intelligence approach for drug discovery.* Molecular diversity, 25(3), 1315–1360. https://doi.org/10.1007/s11030-021-10217-3

[28] Hashimoto, D. A., Witkowski, E., Gao, L., Meireles, O., & Rosman, G. (2020). *Artificial Intelligence in Anesthesiology: Current Techniques, Clinical Applications, and Limitations.* Anesthesiology, 132(2), 379–394. https://doi.org/10.1097/ALN.0000000000002960

[29] Johnson, K. W., Torres Soto, J., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., Ashley, E., & Dudley, J. T. (2018). *Artificial Intelligence in Cardiology.* Journal of the American College of Cardiology, 71(23), 2668–2679. https://doi.org/10.1016/j.jacc.2018.03.521

[30] Jiang, Y., Yang, M., Wang, S., Li, X., & Sun, Y. (2020). *Emerging role of deep learning-based artificial intelligence in tumor pathology.* Cancer communications (London, England), 40(4), 154–166. https://doi.org/10.1002/cac2.12012

[31] van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A., & Viergever, M. A. (2022). *Explainable artificial intelligence (XAI) in deep learning-based medical image analysis.* Medical image analysis, 79, 102470. https://doi.org/10.1016/j.media.2022.102470

[32] Mishra, K., & Leng, T. (2021). *Artificial intelligence and ophthalmic surgery.* Current opinion in ophthalmology, 32(5), 425–430. https://doi.org/10.1097/ICU.0000000000000788

[33] Kazim, E., & Koshiyama, A. S. (2021). *A high-level overview of AI ethics.* Patterns (New York, N.Y.), 2(9), 100314. https://doi.org/10.1016/j.patter.2021.100314

[34] Google: artificial intelligence principles. 2020. https://ai.google/principles/ [14-06-24]

[35] AI Act. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai [14-06-24]

[36] Chauhan, C., & Gullapalli, R. R. (2021). *Ethics of AI in Pathology: Current Paradigms and Emerging Issues.* The American journal of pathology, 191(10), 1673–1683. https://doi.org/10.1016/j.ajpath.2021.06.011

[37] P. Kottarathil. *Polycystic ovary syndrome (PCOS).* 2020. Kaggle dataset. Disponible en: https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos [14-06-24]

[38] *PCOS 2023 Dataset.* Kaggle dataset. Disponible en: https://www.kaggle.com/datasets/sahilkoli04/pcos2023 [14-06-24]

[39] Yu, O., Christ, J. P., Schulze-Rath, R., Covey, J., Kelley, A., Grafton, J., Cronkite, D., Holden, E., Hilpert, J., Sacher, F., Micks, E., & Reed, S. D. (2023). *Incidence, prevalence, and trends in polycystic ovary syndrome diagnosis: a United States population-based study from 2006 to 2019.* American journal of obstetrics and gynecology, 229(1), 39.e1–39.e12. https://doi.org/10.1016/j.ajog.2023.04.010

[40] American Red Cross. *Acerca de la sangre* https://www.redcrossblood.org/espanol/donar-sangre/acerca-de-la-sangre.html#:~:text=El%20grupo%20O%20positivo%20es,relativamente%20alto%20de%20grupo%20B. [14-06-24]

[41] *Gradio.* https://www.gradio.app/ [14-06-24]

Apéndice A

Modelos

Modelos	DB (1)	DB (2)	DB + eco	DB
Decision Tree	96.7	97.6	98.0	97.6
Random Forest	96.7	97.8	99.1	99.1
Gradient Boosting	94.2	94.4	98.9	99.1
XGBoost	96.5	97.4	98.5	98.9
AdaBoost	87.4	90.7	98.1	98.1
SVM	85.3	84.4	92.2	92.0
Neural Network	80.0	85.7	98.0	97.8
Stacking	94.8	95.7	99.6	99.4

Tabla A.1: *Accuracy* de cada modelo utilizado. De izquierda a derecha, se presentan los modelos realizados con los datos básicos del primer *dataset*, los datos básicos del segundo *dataset*, los datos básicos y de ecografía del primer *dataset* y los datos básicos, de ecografía y de analítica del primer *dataset*.

Modelos	DB (1)	DB (2)	DB + eco	DB
Decision Tree	94.9	94.7	96.9	96.3
Random Forest	94.9	95.1	98.6	98.6
Gradient Boosting	91.1	86.9	98.3	98.6
XGBoost	94.6	94.3	97.8	98.3
AdaBoost	79.4	77.7	97.1	97.1
SVM	77.0	64.4	87.9	87.5
Neural Network	63.0	59.8	96.9	96.6
Stacking	91.9	89.8	99.4	99.1

Tabla A.2: *F score* de cada modelo utilizado. De izquierda a derecha, se presentan los modelos realizados con los datos básicos del primer *dataset*, los datos básicos del segundo *dataset*, los datos básicos y de ecografía del primer *dataset* y los datos básicos, de ecografía y de analítica del primer *dataset*.

Modelos	DB (1)	DB (2)	DB + eco	DB
Decision Tree	96.3	97.1	97.9	97.5
Random Forest	96.1	96.9	99.2	99.0
Gradient Boosting	93.2	90.8	98.7	99.0
XGBoost	95.9	97.3	98.8	99.0
AdaBoost	84.0	84.5	97.6	97.7
SVM	82.7	77.0	90.7	90.3
Neural Network	72.8	72.2	97.8	97.5
Stacking	93.7	92.3	99.6	99.3

Tabla A.3: Área bajo la curva ROC de cada modelo utilizado. De izquierda a derecha, se presentan los modelos realizados con los datos básicos del primer *dataset*, los datos básicos del segundo *dataset*, los datos básicos y de ecografía del primer *dataset* y los datos básicos, de ecografía y de analítica del primer *dataset*.