EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Instituto Tecnológico de Orizaba

DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

OPCIÓN I.- TESIS

TRABAJO PROFESIONAL

"Desarrollo de plataforma médica para la salud femenina, relacionada con el síndrome de ovario poliquístico, utilizando técnicas de inteligencia artificial."

QUE PARA OBTENER EL GRADO DE:

MAESTRA EN SISTEMAS COMPUTACIONALES

PRESENTA:

I.S.C Karen Bozziere Solís

DIRECTORA DE TESIS:

MCE Beatriz Alejandra Olivares Zepahua

CODIRECTORA DE TESIS:

Dra. Laura Nely Sánchez Morales

ORIZABA, VERACRUZ, MÉXICO **MARZO 2026**

Autorización de impresión

Educación
Secretaría de Educación Pública

TECNOLÓGICO NACIONAL DE MÉXICO

Instituto Tecnológico de Orizaba
Departamento de División de Estudios de Posgrado e Investigación

Orizaba, Veracruz, 6/marzo/2026
Dependencia: **División de Estudios de Posgrado e Investigación**
Asunto: **Autorización de Impresión**
OPCION: **I**

C. KAREN BOZZIERE SOLIS
CANDIDATO A GRADO DE MAESTRO EN:
SISTEMAS COMPUTACIONALES
P R E S E N T E.-

De acuerdo con el Reglamento de Titulación vigente de los Centros de Enseñanza Técnica Superior, dependiente de la Dirección General de Institutos Tecnológicos de la Secretaría de Educación Pública y habiendo cumplido con todas las indicaciones que la Comisión Revisora le hizo respecto a su Trabajo Profesional titulado:

" Desarrollo de una plataforma médica para la salud femenina, relacionada con el síndrome de ovario poliquístico, utilizando técnicas de inteligencia artificial."

comunico a Usted que este Departamento concede su autorización para que proceda a la impresión del mismo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
CIENCIA – TÉCNICA - CULTURA®

DR. OSCAR BAEZ SENTIES
JEFE DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

Educación
Secretaría de Educación Pública
TECNOLÓGICO NACIONAL DE MÉXICO
INSTITUTO TECNOLÓGICO DE ORIZABA
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

OG-13-F06

2026 año de Margarita Maza

100% LIBRE DE PLÁSTICO

Av. Oriente 9 #852 Colonia Emiliano Zapata, C.P. 94320 Orizaba, Veracruz
Tel. 01 (272) 1105360 info@orizaba.tecnm.mx **www.orizaba.tecnm.mx**

Revisión de trabajo escrito

Educación
Secretaría de Educación Pública

TECNOLÓGICO NACIONAL DE MÉXICO

Instituto Tecnológico de Orizaba
División de Estudios de Posgrado e Investigación

Orizaba, Veracruz, 28/noviembre/2025
Asunto: **Revisión de trabajo escrito**

C. OFELIA LANDETA ESCAMILLA
JEFA DE LA DIVISIÓN DE ESTUDIOS
DE POSGRADO E INVESTIGACIÓN
P R E S E N T E.-

Los que suscriben, miembros del jurado, han realizado la revisión de la Tesis del (la) C.

KAREN BOZZIERE SOLIS

La cual lleva el título de:

Desarrollo de una plataforma médica para la salud femenina, relacionada con el síndrome de ovario poliquístico, utilizando técnicas de Inteligencia Artificial.

Y concluyen que se acepta.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
CIENCIA – TÉCNICA - CULTURA®

PRESIDENTE:	M.C.E. BEATRIZ ALEJANDRA OLIVARES ZEPAHUA	FIRMA
SECRETARIO:	DRA. LAURA NELY SÁNCHEZ MORALES	FIRMA
VOCAL:	M.C. MÓNICA RUIZ MARTÍNEZ	FIRMA
VOCAL SUP.:	DR. JOSÉ LUIS SÁNCHEZ CERVANTES	FIRMA

TA-09-18

2025 Año de La Mujer Indígena

Av. Oriente 9 #852 Colonia Emiliano Zapata, C.P. 94320
Orizaba, Veracruz. Tel. 01 (272) 1105360
e-mail: info@orizaba.tecnm.mx www.orizaba.tecnm.mx

100% LIBRE DE PLÁSTICO

Carta de originalidad

Educación
Secretaría de Educación Pública

TECNOLÓGICO NACIONAL DE MÉXICO

Instituto Tecnológico de Orizaba
División de Estudios de Posgrado e Investigación

Orizaba, Veracruz, 18/septiembre/2025
Asunto: Carta de Originalidad de Tesis

ALFONSO FLORES LEAL
COORDINADOR DEL PROGRAMA DE MAESTRÍA EN SISTEMAS COMPUTACIONALES
PRESENTE

Por este conducto me permito informarle que, con base en los resultados obtenidos por el software institucional, es posible afirmar que el porcentaje de similitud de la tesis **"DESARROLLO DE PLATAFORMA MÉDICA PARA LA SALUD FEMENINA, RELACIONADA CON EL SÍNDROME DE OVARIO POLIQUÍSTICO, UTILIZANDO TÉCNICAS DE INTELIGENCIA ARTIFICIAL"** está dentro de los parámetros establecidos por la División de Estudios de Posgrado e Investigación. Dicha tesis fue desarrollada por la estudiante **KAREN BOZZIERE SOLÍS**, con número de control **M18011163**, dirigida por una servidora, BEATRIZ ALEJANDRA OLIVARES ZEPAHUA, perteneciente al programa de **MAESTRÍA EN SISTEMAS COMPUTACIONALES**.

Por lo descrito anteriormente, se establece que la tesis mencionada **ESTÁ LIBRE DE PLAGIO**.

Sin otro particular, aprovecho la ocasión para enviarle un cordial saludo.

ATENTAMENTE
EXCELENCIA EN EDUCACIÓN TECNOLÓGICA®
Ciencia – Técnica – Cultura ®

BEATRIZ ALEJANDRA OLIVARES ZEPAHUA
Directora de Tesis

Vo.Bo.

JOSÉ LUIS SÁNCHEZ CERVANTES
Presidente del Consejo de la Maestría en Sistemas Computacionales

2025 Año de La Mujer Indígena

Av. Oriente 9 #852 Colonia Emiliano Zapata, C.P. 94320
Orizaba, Veracruz. Tel. 01 (272) 1105360
e-mail: info@orizaba.tecnm.mx www.orizaba.tecnm.mx

100% PLÁSTICO

Carta de cesión de derechos

Declaración de originalidad y cesión de derechos

Orizaba, Veracruz, el día 11 del mes de marzo del año 2026

El(la) que suscribe

C. Karen Bozziere Solis

Declaro que esta tesis, con una extensión de 142 cuartillas, constituye el registro escrito del trabajo titulado:

"Desarrollo de una plataforma médica para la salud femenina, relacionada con el síndrome de ovario poliquístico, utilizando técnicas de inteligencia artificial"

correspondiente al programa: Maestría en Sistemas Computacionales, realizada bajo la asesoría y dirección de la MCE Beatriz Alejandra Olivarez Zepahua y la Dra. Laura Nely Sánchez Morales, y que dicho trabajo ha sido elaborado por mi autoría y no ha sido sometido previamente a evaluación en ninguna otra institución.

Todos los datos y referencias a materiales previamente publicados se encuentran debidamente identificados mediante el crédito correspondiente e incluidos en las notas bibliográficas y citas respectivas. En los casos que así lo requieren, se cuenta con las autorizaciones necesarias de quienes poseen los derechos patrimoniales. En consecuencia, asumo la responsabilidad ante cualquier reclamación o litigio relacionado con derechos de propiedad intelectual, deslindando de toda responsabilidad al Tecnológico Nacional de México campus Orizaba.

Asimismo, declaro que al presentar esta tesis cedo los derechos del trabajo al Tecnológico Nacional de México campus Orizaba para su difusión con fines académicos y de investigación, conforme a las regulaciones propias de la institución. En caso de existir acuerdos de confidencialidad relacionados con la información contenida en el documento, estos se harán constar por escrito para que se omitan las secciones correspondientes.

Los usuarios de la información no deberán reproducir el contenido textual, gráficas o datos de este trabajo sin el permiso expreso del autor y del director del trabajo. Dicho permiso podrá solicitarse por escrito a la institución correspondiente. En caso de otorgarse, el usuario deberá otorgar el reconocimiento adecuado y citar la fuente correspondiente.

Karen Bozziere Solis

Agradecimientos

Al Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT) por el financiamiento otorgado y al Tecnológico Nacional de México (TecNM) por las facilidades proporcionadas para la realización de este trabajo.

Índice General

Índice de Tablas

Índice de Figuras

Índice de Listados

Resumen

El Síndrome de Ovario Poliquístico (SOP) es una condición endocrina relativamente común entre mujeres en edad reproductiva, cuyo diagnóstico temprano es esencial para evitar complicaciones a largo plazo, este diagnóstico presenta desafíos debido a la diversidad de síntomas del síndrome y a la necesidad de pruebas laboratoriales y de gabinete que no siempre es posible realizar, sobre todo en los casos de mujeres con menores recursos económicos y acceso limitado a especialistas médicos. Sin embargo, las técnicas de Inteligencia Artificial (IA) desmotraron su utilidad para identificar patrones complejos en grandes conjuntos de datos, incluidos datos médicos, por lo que se consideran una alternativa viable para proveer sugerencias a médicos generales e incluso a médicos especialistas.

Actualmente, el diagnóstico del SOP se basa en criterios clínicos y pruebas de laboratorio y gabinete, lo que llega a ser costoso y requerir tiempo. Aunque existen algunas aplicaciones de IA en el ámbito específico para el diagnóstico de SOP, el presente proyecto se enfoca en el sector de primer nivel de atención médica (medicina general y familiar), lo que incluye al sector rural o semi rural.

Este proyecto propone el desarrollo de una plataforma de salud basada en IA para el diagnóstico temprano del SOP. La plataforma integra algoritmos de aprendizaje automático entrenados con datos clínicos y biomarcadores relevantes. Además, la plataforma incluye una interfaz de usuario intuitiva para facilitar la entrada de datos de la paciente y apoyar al médico en la entrega rápida y precisa de un diagnóstico, esto mediante una aplicación para dispositivo móvil dirigida a las pacientes y una aplicación web dirigida al personal médico.

Se espera que esta plataforma de salud apoye significativamente la detección del SOP, al complementar los procedimientos tradicionales de diagnóstico como la historia clínica y el examen físico que realiza un médico familiar o general con tecnología de punta, impactando positivamente al sector de mujeres de bajos recursos económicos que no tienen acceso fácil a una consulta de especialidad, lo que permitirá intervenciones tempranas y una mejor gestión de este síndrome.

Abstract

Polycystic Ovary Syndrome (PCOS) is a relatively common endocrine condition among women of reproductive age. Early diagnosis is essential to prevent long-term complications. This diagnosis presents challenges due to the diversity of symptoms and the need for laboratory and office tests, which are not always possible, especially in women with fewer economic resources and limited access to medical specialists. However, Artificial Intelligence (AI) techniques have proven useful in identifying complex patterns in large data sets, including medical data, and are therefore considered a viable alternative for providing suggestions to general practitioners and even specialists.

Currently, PCOS diagnosis is based on clinical criteria and laboratory and office tests, which can be costly and time-consuming. Although there are some AI applications in the specific field for PCOS diagnosis, this project focuses on the primary healthcare sector (general and family medicine), including the rural or semi-rural sector.

This project proposes the development of an AI-based health platform for the early diagnosis of PCOS. The platform integrates machine learning algorithms trained with relevant clinical data and biomarkers. Additionally, the platform includes an intuitive user interface to facilitate patient data entry and support physicians in providing a rapid and accurate diagnosis. This is achieved through a mobile application for patients and a web application for medical staff. This health platform is expected to significantly support PCOS detection by complementing traditional diagnostic procedures such as the medical history and physical examination performed by a family doctor or general practitioner with cutting-edge technology. This will positively impact low-income women who lack easy access to a specialist consultation, enabling early interventions and better management of this syndrome.

Introducción

El acceso a servicios de salud de calidad y la atención oportuna es un derecho humano ampliamente reconocido, sin embargo, su realización plena enfrenta obstáculos significativos, especialmente para las mujeres. Aunque se reconoce este derecho a nivel internacional, su implementación en áreas cruciales como la salud reproductiva y sexual sigue siendo insuficiente.

Dentro de las diversas condiciones de salud que impactan a las mujeres, el Síndrome de Ovario Poliquístico (SOP) es un desafío destacado. Este trastorno, caracterizado por desequilibrios hormonales, presenta un impacto significativo en la calidad de vida si no se aborda de manera adecuada. Sin embargo, los tabúes culturales, la falta de acceso a recursos especializados y la diversidad de síntomas dificultan el diagnóstico temprano y el manejo efectivo de esta condición, lo que a menudo conduce a diagnósticos incorrectos.

Además, la falta de opciones específicas en español para el seguimiento y manejo del SOP en aplicaciones de salud personal representa un vacío importante en el apoyo digital disponible para las mujeres hispanohablantes. Esta brecha digital agrava aún más los desafíos que enfrentan las mujeres en el acceso a la atención médica integral y especializada.

A todo esto, se suma la complejidad del sistema de atención médica en México, que abarca una amplia gama de servicios públicos y privados. Esta diversidad en los sistemas de atención médica presenta desafíos adicionales para abordar eficazmente el SOP en diferentes sectores de la población, destacando la necesidad urgente de medidas integrales que limiten las barreras de acceso y aumenten la conciencia pública sobre esta condición.

El presente documento sigue esta estructura: primero, en el capítulo uno se muestran los antecedentes, centrados en las bases principales para llevar a cabo el proyecto. Luego, en el capítulo dos, se aborda el estado de la práctica, el cual consta de consultar trabajos e investigaciones previas del área, así como la propuesta de solución resultante. El capítulo tres, muestra el desarrollo de la plataforma, mientras que el capítulo cuatro muestra los resultados obtenidos en la ejecución de las aplicaciones que forman la plataforma. Finalmente, se presentan las conclusiones, el trabajo a futuro y la bibliografía.

Capítulo 1 . Antecedentes

El primer capítulo del proyecto de tesis proporciona un entorno integral que contextualiza los elementos principales, como son el marco teórico, situación tecnológica, económica y operativa de la empresa, planteamiento del problema, objetivo general, objetivos específicos y la justificación; mismos que darán paso a la investigación del tema. Esta sección muestra el trasfondo esencial para comprender el proyecto.

1.1 Marco teórico

Este apartado sitúa los conceptos de investigación dentro del contexto académico, con el fin de mostrar un espectro más amplio dentro del campo de estudio.

1.1.1 Salud de la mujer

De acuerdo a la Enciclopedia Médica en línea del gobierno norteamericano, el término "salud de las mujeres" se refiere a la rama de la Medicina que se enfoca en el tratamiento y diagnóstico de enfermedades y padecimientos que afectan el bienestar físico y emocional de una mujer [1].

1.1.1.1 Síndrome de ovario poliquístico

El síndrome de ovario poliquístico (SOP), es una disfunción endocrino-metabólica con una prevalencia entre el 6 y el 10% en mujeres de edad reproductiva. Se caracteriza por hiperandrogenismo, oligo-anovulación y morfología de ovario poliquística. Representa la causa más frecuente de infertilidad anovulatoria e hiperandrogenismo en mujeres. Además, la mayoría de las pacientes con SOP desarrollan resistencia insulínica periférica (RI) y poseen un mayor riesgo de desarrollar diabetes mellitus tipo 2 (DM2).

Su etiología es multifactorial y es tema de investigación relevante ya que el SOP se considera un desorden genético complejo, en donde numerosos genes contribuyen parcialmente al fenotipo sin que alguno de ellos constituya uno de riesgo mayor. El fenotipo es también influenciado por factores ambientales, siendo finalmente la suma de ambos factores la que dispara el síndrome. Diversos autores sugieren que existen factores genéticos implicados en el desarrollo del síndrome, debido a que se ha observado un fuerte componente de agregación

familiar, siendo afectadas entre el 20 y el 40% de las parientes de primer grado de las mujeres con SOP. En los últimos años se ha propuesto la teoría de que la programación fetal también sea responsable de la patogenia del SOP; bajo esta teoría, la exposición prenatal a andrógenos (EPA) es uno de los principales candidatos como factor reprogramador. Independiente de la fuente del exceso de andrógenos prenatal, estas observaciones sugieren que el SOP estaría programado desde el útero por la influencia del aumento de andrógenos e insulina, posiblemente mediante la alteración de la expresión génica durante la vida pre y postnatal [2].

Más allá de las incomodidades que presenta el SOP, su mayor peligro implica que se derive en infertilidad, cáncer endometrial, diabetes, síndrome metabólico, enfermedades cardiacas y esteatohepatitis no alcohólica.

El SOP no tiene cura, pero sí tratamiento para controlar los síntomas y lidiar con algunas condiciones derivadas como la infertilidad.

1.1.1.2 Sintomatología y Diagnóstico

Las manifestaciones clínicas de SOP son muy diversas, las más destacadas son ciclos menstruales irregulares, dolor pélvico, afectaciones en el peso, exceso de vello corporal y/o facial, acné, manchas oscuras en la piel, alopecia, entre otros. Varios de estos síntomas son comunes con otras enfermedades y esto, aunado a tabúes culturales, resulta intimidante para quienes lo padecen SOP, sobre todo en zonas con difícil acceso a servicios de salud, además de que la similitud de síntomas lleva a diagnósticos equivocados con relativa facilidad. Se tienen información anecdótica de casos en que las pacientes fueron derivadas por el médico familiar a Dermatología (debido al acné) o a Medicina Interna (por la resistencia a la insulina) o a Nutrición (por el sobrepeso) cuando en realidad se trata de síntomas de SOP y no de otras enfermedades.

Ante la sospecha de SOP, los (as) médicos (as) especialistas ordenan primeramente estudios laboratoriales que incluyen tanto elementos tradicionales como hormonales, que les permiten descartar otro tipo de enfermedades, estos estudios tienen un costo aproximado de dos mil pesos. Generalmente, para confirmar el diagnóstico de SOP, se necesitan también pruebas de gabinete como la ecografía cuyo costo aproximado es de mil quinientos pesos. Estos costos resultan prohibitivos para las pacientes de escasos recursos.

1.1.1.2.1 Hiperandrogenismo

El hiperandrogenismo (hirsutismo, pérdida de cabello de patrón femenino y acné) es una característica diagnóstica clave del síndrome de ovario poliquístico, se presenta entre el 60% y el 100% de los casos, bien sea clínico (manifestaciones físicas) o bioquímico (presente en las alteraciones hormonales que se detecta en pruebas laboratoriales). El hiperandrogenismo es difícil de evaluar, con variaciones según los métodos, el origen étnico y los factores de confusión, incluido el exceso de peso y la etapa de la vida. La evaluación del hiperandrogenismo bioquímico ha sido controvertida en términos del tipo óptimo de andrógenos y ensayos, la definición de rangos normales, las superposiciones entre los controles y el síndrome de ovario poliquístico y las cuestiones de acceso y costo para los ensayos de alta calidad.

Los signos y síntomas de un exceso grave de andrógenos provocan virilización (p. ej., calvicie de patrón masculino, hirsutismo grave y clitoromegalia) y masculinización, que son poco frecuentes en el síndrome de ovario poliquístico. La evidencia clínica de exceso de andrógenos leve a moderado es más común, incluido el hirsutismo, el acné y la caída del cabello de patrón femenino. Las interrelaciones de estas características clínicas siguen sin estar claras, con variaciones según el origen étnico, y requieren capacitación clínica, vigilancia y habilidad para evaluarlas. Estas características tienen un impacto considerable en la calidad de vida de las mujeres con síndrome de ovario poliquístico y la carga del tratamiento, incluida la terapia cosmética, es significativa.

La evaluación del hiperandrogenismo clínico es de bajo costo, en relación con las evaluaciones bioquímicas del hiperandrogenismo y es de mayor relevancia para las personas afectadas por el síndrome de ovario poliquístico; sin embargo, tiene un componente subjetivo, por lo que se requiere que el personal médico utilice una evaluación estandarizada que permita minimizar este componente [3].

Algunos de los síntomas del hiperandrogenismo clínico, como el acné, también se presentan en mujeres sanas durante la pubertad y adolescencia, por lo que existen escalas de valoración, como es las escalas de Ferriman-Gallwey para evaluar hirsutismo, la escala de Ludwig para evaluar la alopecia androgénica femenina, evaluaciones clínicas de acné, entre otras.

En el caso de la perdida de cabello, en este ámbito denominada alopecia androgénica femenina (FAGA) típica comienza con una específica “pérdida difusa de cabellos de las regiones parietal y frontovertical respetando la línea de implantación frontal”. Ludwig llamó a este proceso “rarefacción”. En la clasificación de la escala de Ludwig se describieron tres grados o tipos progresivos de FAGA. Grado I o mínimo, grado II o moderado y grado III o intenso [4].

1.1.1.2.1.1 Hirsutismo

El hirsutismo es un exceso de vello corporal generalmente en lugares donde no es común; el hirsutismo facial es un síntoma angustiante de hiperandrogenismo en mujeres con síndrome de ovario poliquístico de cualquier edad.

El crecimiento excesivo del vello corporal tiene un impacto negativo en el bienestar emocional y la calidad de vida. El tratamiento cosmético y los AOC (Anticonceptivos Orales Combinados) generalmente se consideran tratamientos de primera línea para el hirsutismo en mujeres, incluso en el síndrome de ovario poliquístico [3].

El método más común para calificar el crecimiento del vello terminal facial y corporal es la escala Ferriman y Gallwey (FG) modificada. Es considerado como el estándar de oro para la evaluación de hirsutismo clínico[5].

1.1.1.2.2 Anovulación

La disfunción ovulatoria es una característica diagnóstica clave del SOP y se distingue por ciclos menstruales irregulares que reflejan dicha disfunción, como quedó descrito en los criterios médicos internacionales de Rotterdam de 2003 [6] y en la Guía Internacional de SOP de 2018[7]. Sin embargo, la disfunción ovulatoria también ocurre con ciclos regulares; por lo cual, si hay sospecha de SOP y los ciclos son regulares, es necesario confirmar la anovulación mediante la evaluación hormonal.

Es necesario aclarar que los ciclos irregulares y la disfunción ovulatoria también son un componente normal de las transiciones puberales y menopáusicas y, por lo tanto, definir la anomalía en estas etapas de la vida sigue siendo un desafío. De hecho, la mayor controversia en cuanto al criterio diagnóstico del SOP se produce durante la transición puberal. La maduración fisiológica del eje hipotalámico-pituitario ovárico se produce a lo largo de los años y la

ovulación y los ciclos en las adolescentes no coinciden con los de las mujeres en edad reproductiva [3].

1.1.2 Inteligencia artificial (IA)

Se considera una rama de la ingeniería computacional que implementa conceptos y soluciones novedosas para resolver desafíos complejos bajo la idea de emular la inteligencia humana; cuenta con avances continuos en velocidad, capacidad y programación de software electrónicos. En el ámbito médico en particular, el término se aplica a una amplia gama de elementos como la robótica, el diagnóstico médico, las estadísticas médicas entre otras. La IA en la Medicina tiene dos ramas principales: virtual y física. La rama virtual incluye enfoques informáticos que van desde la gestión de la información mediante aprendizaje profundo hasta el control de los sistemas de gestión de la salud, incluidos los registros médicos electrónicos, y la orientación activa de los(as) médicos(as) en sus decisiones de tratamiento, considerando que la decisión final siempre es del personal médico. La rama física está representada por robots que se utilizan para ayudar al paciente anciano o al cirujano que lo atiende. También están incorporados en esta rama los nano robots dirigidos, un nuevo sistema único de administración de fármacos [8].

1.1.2.1 Minería de datos

La minería de datos es una actividad relacionada con la recopilación de datos, el uso de datos históricos para descubrir conocimiento, información, regularidades, patrones o relaciones en datos. El resultado de la minería de datos se utiliza como apoyo en la toma de decisiones.

La minería de datos no es un campo de la ciencia independiente, sino que está estrechamente relacionada con otros como las bases de datos, las estadísticas, la recuperación de información y la inteligencia artificial.

1. Base de datos - Minería de datos: La recopilación de datos utilizados en la minería de datos proviene de una base de datos. Los datos que se extraen, a partir de los consejos del especialista, se separan de los datos operativos en la base de datos.
2. Estadísticas - Minería de datos: en la toma de decisiones, las estadísticas requieren datos a partir de la recopilación de datos, el muestreo de datos y la probabilidad. La minería

de datos implica determinar muestras de datos, analizar y presentar resultados utilizando técnicas estadísticas.

3. Búsqueda de información - Minería de datos: La búsqueda de información es una de las actividades del proceso de minería de datos que incluye interpretación, análisis y almacenamiento de datos.
4. Inteligencia artificial - Minería de datos: una rama de la inteligencia artificial es el aprendizaje automático. El aprendizaje automático es una disciplina científica importante en la minería de datos donde los sistemas informáticos aprenden de los datos de entrenamiento utilizados.

La minería de datos permite trabajar diversas tareas, de acuerdo al objetivo que se persiga en la obtención de conocimiento. Las principales tareas son;

1. Descripción

   El objetivo de las tareas descriptivas es identificar patrones que aparecen con frecuencia y convertir estos patrones en reglas que describen alguna situación particular de negocio; por ejemplo, para explicar qué características tienen en común los clientes de una empresa.
2. Clasificación

   Permite identificar los patrones que relacionan las variables con etiquetas o clases conocidas, para posteriormente encontrar la etiqueta de un nuevo elemento. Es una de las tareas de más uso en el área de salud, parte de la experiencia del personal médico para identificar si una persona es susceptible de padecer una determinada enfermedad o no, en este caso las clases serían “Sí” y “No” respecto a la variable “Está enfermo”. También hay casos donde las clases no son binarias sino múltiples (por ejemplo: Sin Enfermedad, Diabetes Tipo I y Diabetes Tipo II).
3. Predicción

   El propósito de la predicción es semejante a la clasificación, con la diferencia de que, en la predicción, no existen etiquetas o clases, sino valores continuos.
4. Agrupación

   Identifica grupos de elementos que tienen valores similares. Las formas de datos que es posible agrupar son el resultado de observaciones, registros de datos o clases y objetos

que tienen similitudes. A diferencia de la clasificación y la predicción, no hay elementos para comparar el resultado de la agrupación.

5. Asociación

   Es una búsqueda de atributos que siempre aparecen al mismo tiempo. La probabilidad de que los atributos aparezcan simultáneamente se mide mediante valores de confianza.

Las tareas de clasificación y predicción se engloban bajo el concepto de "aprendizaje supervisado", debido a que en el conjunto de datos original existen tanto las variables como las etiquetas o valores a predecir, por lo que es posible comparar los resultados de algoritmos con la información conocida. Por otro lado, la agrupación, la descripción y la asociación se conocen como "aprendizaje no supervisado" porque no se conocen de antemano los grupos o reglas a encontrar, por lo que no se cuenta con elementos para comparar los resultados [9].

1.1.2.2 KDD

Knowledge Discovery in Databases (KDD, Descubrimiento de Conocimiento en Bases de Datos) es el proceso de descubrimiento automático de patrones, reglas y otros contenidos regulares previamente desconocidos que están presentes de manera implícita en grandes volúmenes de datos. La minería de datos (MD) denota el descubrimiento de patrones en un conjunto de datos previamente preparado de una manera específica. MD se utiliza a menudo como sinónimo de KDD. Sin embargo, estrictamente hablando, MD es sólo una fase de todo el proceso de KDD [10].

Si bien las variantes de KDD varían de 5 a 7 pasos, la mayoría de los autores coinciden con los siguientes 5 pasos:

- Selección de variables. A partir de una base o conjunto de datos, se determinan los datos objetivo y se eligen las variables que se utilizarán para evaluar el descubrimiento de conocimiento con la ayuda de expertos del negocio.
- Preprocesamiento. Esta etapa se trata de adaptar los datos con los que se trabaja e incorpora el concepto de limpieza de datos. Esto es debido a que las bases de datos operativas, es decir, las que provienen del trabajo diario, tienen condiciones distintas a las requeridas por la MD; por ejemplo, en los registros médicos que se recaban en las

salas de urgencias, es válido que muchos de los datos queden vacíos, sin embargo son datos importantes para el objetivo de conocimiento determinado en la fase anterior. Por lo tanto, se establecen decisiones respecto al tratamiento que se dará a los registros con datos incompletos, no confiables y/o defectuosos. El tratamiento va desde la eliminación de los registros para fines de la MD hasta la ejecución de algoritmos para predecir valores faltantes.

- Transformación. Esta fase se concentra en ajustar los datos preprocesados para facilitar su uso en los algoritmos de MD. Esto se hace reduciendo el alcance en términos de variedad; por ejemplo, se discretizan valores continuos, se agrupan en rangos de valores o se normalizan.
- Aplicación de Algoritmos de Minería de Datos. La fase se centra en examinar los datos transformados para buscar patrones de interés mediante el uso de algún algoritmo de acuerdo al tipo de objetivo elegido en la primera fase; por ejemplo, algoritmos de clasificación, de agrupación, de asociación, entre otros.
- Interpretación/Evaluación. En la fase final es aquella durante la cual los resultados se entregan para su interpretación y validación por parte de los expertos del negocio. Idealmente, se esperan representaciones gráficas para facilitar la interpretación y análisis [11].

1.1.2.2.1 SMOTE

Hay casos en las tareas de clasificación en que se dice que las clases están “desequilibradas” o “desbalanceadas”, es decir, en los datos las clases no se distribuyen de manera uniforme. Por ejemplo, en los casos de salud, para apoyar en el diagnóstico de una enfermedad como el SOP, donde el 10% de la población femenina en edad fértil la presenta, las clases serán desequilibradas, debido a que de 10 registros de mujeres sólo 1 tendrá el valor SÍ en la variable “Tiene SOP” y 9 tendrán el valor NO en dicha variable.

Los algoritmos para clasificación generalmente no trabajan bien con clases desequilibradas, por lo que se hace necesario “ajustar” los datos.

SMOTE es un algoritmo de sobre muestreo en el que la clase minoritaria aumenta sus muestras mediante la creación de ejemplos "sintéticos" en lugar de sobre muestreo con reemplazo siempre

que las características de la muestra sean valores continuos. Esto para abordar el problema de desequilibrio de clases en el aprendizaje supervisado, especialmente cuando la clase minoritaria es significativamente menor en cantidad que la clase mayoritaria. Se generan ejemplos sintéticos de una manera menos específica de la aplicación, operando en el "espacio de características" en lugar del "espacio de datos". La clase minoritaria se sobre muestrea tomando cada muestra de clase minoritaria e introduciendo ejemplos sintéticos a lo largo de los segmentos de línea que unen a cualquiera o todos los vecinos más cercanos de la clase minoritaria. Dependiendo de la cantidad de sobre muestreo requerida, los vecinos de los vecinos más cercanos se eligen aleatoriamente. Las muestras sintéticas se generan de la siguiente manera: se toma la diferencia entre el vector de características (muestra) en consideración y su vecino más cercano. Se multiplica esta diferencia por un número aleatorio entre 0 y 1 y se suma al vector de características en consideración. Esto provoca la selección de un punto aleatorio a lo largo del segmento de línea entre dos características específicas. Este enfoque obliga efectivamente a que la región de decisión de la clase minoritaria se vuelva más general [12].

1.1.2.2.2 SMOTE-NC

El enfoque de sobre muestreo sintético de minorías - Continuo nominal (SMOTE-NC), es una variante de SMOTE, la cual se utiliza para manejar conjuntos de datos mixtos de características nominales y continuas.

El algoritmo SMOTE-NC se describe a continuación.

1. Cálculo de la mediana: usa la mediana de las desviaciones estándar de todas las características continuas para la clase minoritaria. Si las características nominales difieren entre una muestra y sus vecinos potenciales más cercanos, entonces esta mediana se incluye en el cálculo de la distancia euclidiana. Se utiliza la mediana para penalizar la diferencia de las características nominales por una cantidad que está relacionada con la diferencia típica en los valores de las características continuas.
2. Cálculo del vecino más cercano: utiliza la distancia euclidiana entre el vector de características para el cual se están identificando los k vecinos más cercanos (muestra de clase minoritaria) y los otros vectores de características (muestras de clase minoritaria) utilizando el espacio de características continuo. Para cada característica

nominal diferente entre el vector de características considerado y su vecino potencial más cercano, se incluye la mediana de las desviaciones estándar calculadas previamente en el cálculo de la distancia euclidiana.

3. Generación de la muestra sintética: las características continuas de la nueva muestra sintética de la clase minoritaria se crean utilizando el mismo enfoque de SMOTE. A la característica nominal se le asigna el valor que se presenta en la mayoría de los k vecinos más cercanos [12].

1.1.2.3 Modelo de clasificación

El modelo de clasificación no es otra cosa que el descubrimiento de una función de aprendizaje que clasifica un elemento de datos en una de varias clases predefinidas [13].

Es decir, un modelo de clasificación es el resultado de un algoritmo de minería de datos que, a partir de un conjunto suficientemente amplio, identifica los patrones de relación para "saber" bajo qué condiciones un elemento se clasifica en una u otra clase.

Este tipo de modelos se aplica, por ejemplo, para saber si un equipo electrónico va a fallar o no, si un paciente padece o no una enfermedad, si un paciente ya diagnosticado con una enfermedad se encuentra en una etapa específica de dicha enfermedad, entre otros casos.

1.1.2.4 Naïve Bayes

El clasificador Naïve Bayes es un algoritmo de aprendizaje automático supervisado que se utiliza para tareas de clasificación, es decir, para identificar el valor desconocido de un registro a partir de un conjunto de datos bien conocido. Este algoritmo se basa en teoremas de probabilidad para realizar tareas de clasificación, específicamente en el Teorema de Bayes [14].

1.1.2.5 Perceptrón Multicapas (Multilayer Perceptron)

Es una red neuronal con múltiples capas interconectadas que permite capturar relaciones no lineales complejas entre las variables del problema, donde cada capa contiene un conjunto de elementos perceptuales conocidos como neuronas [15].

1.1.2.6 IBK

Este algoritmo está basado en el principio de los k vecinos más cercanos, el cual predice la clase de una instancia según las distancias relativas a los ejemplos conocidos, ajustando el parámetro K mediante validación cruzada [16].

1.1.2.7 J48

Es un algoritmo de aprendizaje automático utilizado en minería de datos para crear árboles de decisión. Utiliza un conjunto de datos de entrenamiento para construir un árbol que permite tomar decisiones basadas en reglas. Cada nodo del árbol representa una característica de los datos y se ramifica según criterios que maximizan la homogeneidad de las clases [17].

1.1.2.8 Random Forest

Entre los algoritmos para realizar clasificación, uno de los más conocidos es el de árboles de decisión, utiliza una estructura jerárquica (de árbol) para organizar las decisiones que se basan en las características de los datos. Este algoritmo funciona bien con distintos tipos de datos, tanto continuos como discretos y tiene la ventaja de que la representación jerárquica resulta fácil de entender para los usuarios finales, independientemente de cómo se llegó a la jerarquía particular.

Existen diversas variantes del algoritmo, entre ellas "*Random Forest*", que construye múltiples árboles de decisión sobre subconjuntos aleatorios del conjunto de datos, promediando sus resultados para mejorar la precisión y reducir el sobreajuste [18].

1.1.2.9 Instancias clasificadas correctamente

Las instancias clasificadas correctamente son las observaciones que un modelo de clasificación identifica de manera precisa según su clase real. La métrica asociada, conocida como exactitud (*accuracy*), se define como el número de clasificaciones correctamente realizadas por el clasificador dividido entre el número total de clasificaciones. También se conoce como la fracción de instancias clasificadas correctamente [19] y se calcula de la siguiente forma:

$$Exactitud = \frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{VP} + \mathrm{VN} + \mathrm{FP} + \mathrm{FN}}$$

Donde:

VP = Verdaderos positivos (se clasifican como positivos y realmente lo son).

VN = Verdaderos negativos (se clasifican como negativos y realmente lo son).

FP = Falsos positivos (se clasifican como positivos, pero realmente no lo son).

FN = Falsos negativos (se clasifican como negativos, pero realmente no lo son).

1.1.2.10 Kappa Statistic

El coeficiente Kappa de Cohen o *Kappa Statistic* (κ) es una medida estadística que sopesa la concordancia entre anotadores. Generalmente, se considera una medida más robusta que el simple cálculo del porcentaje de exactitud, ya que κ considera que la exactitud es posible que se produzca por casualidad. La estadística kappa tiene un dominio de valores entre - 1 y 1. El máximo valor significa acuerdo completo; cero significa una exactitud al azar y el valor mínimo implica total discordancia (no tiene exactitud) [20].

1.1.2.11 *Mean Absolute Error* (MAE)

El error absoluto medio (MAE) es una medida de la diferencia entre los valores predichos y los reales cuando se promedian todos los valores. El MAE se utiliza habitualmente en el análisis de regresión y en las clasificaciones para comprender el error de predicción del modelo. Si hay una regresión lineal, el MAE representa la distancia promedio desde una línea predicha hasta el valor real. El MAE se define como la suma de los errores absolutos dividida por la cantidad de observaciones. Los valores van de 0 a infinito, y los números más pequeños indican un mejor ajuste del modelo a los datos [21].

1.1.2.12 *Root Mean Squared Error* (RMSE)

La raíz del error cuadrático medio (RMSE) mide la raíz cuadrada de la diferencia cuadrática entre los valores pronosticados y los reales, y se promedia sobre todos los valores. Se utiliza en el análisis de regresión y en las clasificaciones para comprender el error de predicción del modelo. Es una métrica importante para indicar la presencia de valores atípicos y errores de modelo grandes. Los valores van desde cero (0) hasta infinito, y los números más pequeños

indican el modelo que se ajusta mejor a los datos. El RMSE depende de la escala y no debe usarse para comparar conjuntos de datos de diferentes tamaños [21].

1.1.2.13 Precisión

La precisión (*precision*) mide el rendimiento de un algoritmo al clasificar los verdaderos positivos (TP) de entre todos los positivos que identifica. Se define como Precisión = TP/(TP+FP); los valores van de cero (0) a uno (1) y se utiliza en la clasificación binaria. La precisión es una métrica importante cuando el costo de un falso positivo es elevado. Por ejemplo, el costo de un falso positivo es muy elevado si el sistema de seguridad de un avión se equivoca al decir que es seguro volar. Un falso positivo (FP) refleja una predicción positiva que, en realidad, es negativa en los datos [21].

1.1.2.14 Sensibilidad o *Recall*

La sensibilidad mide el rendimiento de un algoritmo al momento de predecir/clasificar correctamente todos los positivos verdaderos (VP) de un conjunto de datos. La sensibilidad se calcula de la siguiente manera:

$$Sensibilidad = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}}$$

Se obtienen valores que van de 0 a 1. Las puntuaciones más altas reflejan una mejor capacidad del modelo para predecir los verdaderos positivos (VP) en los datos. Se utiliza en la clasificación binaria [21]; una alta sensibilidad es importante en casos donde es más grave considerar que no se tiene una enfermedad, por ejemplo, cuando el tratamiento tiene pocos efectos secundarios.

1.1.2.15 Especificidad

Esta medida calcula cuántos negativos reales fueron identificados correctamente, también se utiliza en clasificación binaria; una alta especificidad es importante en los casos donde es grave

diagnosticar como existente una enfermedad y no tenerla, por ejemplo, cuando el tratamiento tiene más efectos indeseables o genera algún tipo de discriminación. Se calcula como:

$$Especificidad = \frac{\mathrm{VN}}{\mathrm{VN} + \mathrm{FP}}$$

Los valores obtenidos también van del cero al uno.

1.1.2.16 F-measure

La medida F se considera como un compromiso entre precisión y sensibilidad. Es alta sólo cuando tanto la sensibilidad como la precisión son altas. Es equivalente a la sensibilidad cuando $\alpha = 0$ y a la precisión cuando $\alpha = 1$. La medida F asume valores en el intervalo [0,1]. Esta medida es relevante cuando se tienen muchos casos de clases negativas (por ejemplo, en los casos de salud donde la mayoría de las muestras son de gente sana) [22].

1.1.2.17 Selección de variables

En aprendizaje automático y estadística, la selección de características, también conocida como selección de variables, selección de atributos o selección de subconjuntos de variables, es el proceso de reducir el número de variables de entrada para un modelo de clasificación. Las técnicas de selección de características se utilizan por varias razones:

- Reducen la complejidad del modelo al descartar algunas características irrelevantes.
- Ayudan al algoritmo de aprendizaje automático a entrenar un modelo más rápido.
- Evitan el sobreajuste [23].

En los siguientes apartados se explican brevemente algunos de los algoritmos implementados dentro de la herramienta Weka para la tarea de selección de variables.

1.1.2.18 CfsSubsetEval

Evalúa el valor de un subconjunto de atributos considerando la capacidad predictiva individual de cada característica, así como el grado de redundancia entre ellas. Se prefieren los subconjuntos de características con alta correlación con la clase y baja inter correlación [24].

1.1.2.19 Best First

Busca en el espacio de subconjuntos de atributos mediante un método de escalada voraz, complementado con una función de retroceso. *Best First* tiene tres posibles inicios: a) comienza con el conjunto vacío de atributos y busca hacia adelante, b) inicia con el conjunto completo de atributos y busca hacia atrás, c) comienza en cualquier punto y busca en ambas direcciones considerando todas las posibles adiciones y eliminaciones de atributos individuales en un punto dado [25].

1.1.2.20 Chi-Cuadrado

La distribución Chi-Cuadrado, también denominada distribución χ^2, es una distribución de probabilidad continua ampliamente utilizada en pruebas de hipótesis estadísticas, en particular en el contexto de pruebas de bondad de ajuste y pruebas de independencia en tablas de contingencia. Se presenta cuando la suma de los cuadrados de las variables aleatorias normales estándar independientes sigue esta distribución [26].

1.1.2.21 InfoGainAttributeEval

Evalúa el valor de un atributo midiendo la ganancia de información con respecto a la clase y permite elegir aquellos atributos que tienen la mayor ganancia [27].

1.1.2.22 ReliefFAttributeEval

Evalúa el valor de un atributo realizando muestras repetidamente sobre una instancia y considerando el valor del atributo dado para la instancia más cercana de la misma clase y de una diferente. Trabaja con datos discretos y continuos [28].

1.1.3 SCRUM

Scrum es un marco de administración que los equipos utilizan para organizarse por cuenta propia y trabajar en aras de alcanzar un objetivo común. Describe un conjunto de reuniones, herramientas y funciones para entregar proyectos de forma eficiente. Al igual que un equipo deportivo que practica para un importante partido, las prácticas de Scrum permiten a los equipos de trabajo gestionarse por cuenta propia, aprender a partir de la experiencia y adaptarse al

cambio. Los equipos de software utilizan Scrum para resolver problemas complejos de forma rentable y sostenible [29].

Si bien se trata de un marco de trabajo dirigido a equipos, dado que se basa en procesos iterativos de desarrollo, es posible adaptarlo a trabajos individuales, tomando de Scrum ideas como iteraciones (*sprints*) y pila de producto (características a cumplir).

1.1.4 Plataforma digital

Una plataforma digital es el software y la tecnología que se utilizan para unificar y optimizar las operaciones de negocio y los sistemas de TI. Una plataforma digital funciona como la columna vertebral de una organización para las operaciones y el compromiso del cliente [30].

En general, una plataforma digital permite estandarizar las operaciones y presentar un espacio unificado para realizar diversas tareas de una organización.

1.1.5 Aplicación para dispositivo móvil

Una aplicación para dispositivo móvil, también conocida como app, es una aplicación que se diseña específicamente para las características de un dispositivo como teléfono celular o tableta, en contraposición de una aplicación diseñada para ejecutarse en una computadora personal (de escritorio o portátil).

Las aplicaciones para dispositivos móviles toman en cuenta las capacidades limitadas de cómputo que manejan este tipo de dispositivos (en comparación con una computadora personal) y aprovechan las ventajas de su ubicuidad y portabilidad, así como su popularidad entre el público, buscando el equilibrio entre ambos [31].

Para la instalación de la app en el dispositivo, generalmente es necesario descargarla de una tienda oficial, como Google Play o Apple Store, dependiendo del sistema operativo que maneje el dispositivo; esto permite un cierto nivel de confianza debido a que estas tiendas realizan una evaluación previa de las aplicaciones antes de permitir que se distribuyan a través de ellas.

1.1.6 Expediente clínico

De acuerdo a la NOM-004-SSA3-2012, el expediente clínico se define como el conjunto único de información y datos personales de un(a) paciente, que se integra dentro de todo tipo de

establecimiento para la atención médica, ya sea público, social o privado; consta de documentos escritos, gráficos, imagenológicos, electrónicos, magnéticos, electromagnéticos, ópticos, magneto-ópticos y de cualquier otra índole, en los cuales, el personal de salud realizará los registros, anotaciones, en su caso, constancias y certificaciones correspondientes a su intervención en la atención médica del paciente, con apego a las disposiciones jurídicas aplicables [32].

Esta norma, junto con la NOM-004-SSA3-2012 (sobre el expediente clínico electrónico), define el contenido y la seguridad esperada de los datos personales y médicos de un(a) paciente, los requisitos que debe cumplir la información y los sistemas que la soporten, en su caso.

1.1.7 Datos personales

De acuerdo a la normativa mexicana, los datos personales son "la información concerniente a una persona física identificada o identificable" [33]. Abarcan una amplia gama de detalles que, de conocerse, permiten revelar la identidad de una persona. Esto no incluye sólo datos básicos como el nombre, la dirección y el número de teléfono, sino también información más específica como la fecha de nacimiento, el número de identificación personal, la dirección de correo electrónico y hasta detalles biométricos como huellas dactilares o características faciales.

Cualquier aplicación computacional que recabe y/o almacene datos personales obliga a sus creadores a proteger la privacidad de estos datos y respetar las leyes y regulaciones relacionadas con su recopilación, almacenamiento y uso. Esto tiene varios impactos directos en la funcionalidad de un sistema computacional de este tipo: de primera instancia, el sistema debe contar con un aviso de privacidad visible, claro y legal; además, debe incluir la funcionalidad necesaria para cumplir con los derechos ARCO (Acceso, Rectificación, Cancelación y Oposición) descritos en la normativa mexicana.

1.1.8 Datos personales sensibles

De acuerdo a la normativa mexicana, los datos personales sensibles son aquellos "que se refieren a la esfera más íntima de su titular, o cuya utilización indebida da origen a discriminación o conlleva un riesgo grave para éste" [33]. En este caso se encuentran los datos médicos, por lo que cualquier aplicación que los contemple contará con los mecanismos de seguridad y

protección especial que exige la ley. Entre los datos médicos se encuentran los datos clínicos que se refieren a elementos que evalúa la paciente o el médico, por ejemplo, la caída del cabello, crecimiento del vello corporal, peso o presión arterial.

1.1.9 Datos laboratoriales

Dentro de un expediente clínico, los datos laboratoriales se refieren a los resultados obtenidos en las diversas pruebas que se realizan en un laboratorio clínico, Las pruebas de laboratorio examinan muestras de sangre, orina o tejidos corporales [34].

Para efectos médicos, las pruebas laboratoriales son distintas de las pruebas de gabinete (radiografía, ecografía, ultrasonido, resonancia magnética, tomografía, entre otras); sin embargo, la normativa sobre Expediente Clínico Electrónico en México no hace esa distinción.

1.2 Situación tecnológica, económica y operativa de la empresa.

En México, el sistema de salud cuenta con tres niveles

1. Primer nivel: Atención primaria o básica, proporcionada en centros de salud, unidades médicas rurales y consultorios médicos. Aquí se brindan servicios de prevención, promoción de la salud, atención médica general y algunas especialidades básicas.

2. Segundo nivel: Atención especializada, ofrecida en hospitales generales y unidades médicas especializadas. En este nivel se realizan diagnósticos y tratamientos más complejos, incluyendo cirugías, atención de urgencias, hospitalización y algunas especialidades médicas y quirúrgicas.

3. Tercer nivel: Atención de alta especialidad, concentrada en hospitales de referencia a nivel estatal o nacional. Este nivel se enfoca en la atención de enfermedades poco comunes o altamente especializadas, así como en la investigación médica y la formación de especialistas.

El entorno hacia el cual va dirigido este proyecto corresponde a la atención de primer nivel, que incluye el sector rural o semi rural, es decir, va dirigido al primer contacto médico que realiza una paciente.

Para el presente proyecto, dentro de dicho nivel de atención se cuenta con el apoyo de la Médica General Guadalupe Verónica Islas Vargas con cédula profesional 6474743, quien dirige un consultorio médico en el municipio de Tlaltetela, Veracruz, México. De acuerdo con el Informe Anual sobre la Situación de Pobreza y Rezago Social 2022 [35], esta comunidad tiene una población de 16,485 habitantes, de la cual el 49.5% son mujeres, tiene un grado de marginación y de rezago social medio, sólo el 24.1% de la población tiene acceso a servicios de salud, el 30% de la población presenta rezago educativo; aunque se indica que el 89% de la población tiene acceso a la seguridad social, no se tiene registro de unidades médicas en el municipio, por lo que se asume que este rubro se refiere a los apoyos federales individuales.

Por otra parte, para efectos de validar la información médica especializada, es decir, la que se obtiene en atención de segundo nivel, para este proyecto se cuenta con la asesoría de la ginecóloga Miriam Rosas Luna con cédula profesional 11741283.

1.3 Planteamiento del problema

El acceso a servicios de salud de calidad y de forma oportuna, es un derecho humano reconocido a nivel internacional; sin embargo, en la práctica, no siempre es posible ejercer este derecho, sobre todo en el caso de las mujeres y, en particular, cuando se trata de salud reproductiva y sexual, al grado de existir un "Día Internacional de Acción por la Salud de la Mujer" para visibilizar esta problemática, instituido desde 1987 [36]. De acuerdo a investigaciones publicadas en 2014, por el entonces CONACYT [37], en muchos países los temas de salud femenina se limitaban al cuidado del embarazo y puerperio, si bien se incorporaron los relacionados con la prevención del cáncer de mama y del cáncer cervicouterino, todavía quedan fuera muchos otros temas igualmente importantes.

En este sentido, el Síndrome de Ovario Poliquístico (SOP), es una enfermedad relacionada con desequilibrios hormonales que se presenta en mujeres en edad fértil, con una sintomatología que llega a afectar la vida diaria de las pacientes y que, si no se trata adecuadamente, es posible que derive en infertilidad, cáncer endometrial, diabetes y esteatohepatitis no alcohólica [38]; esta

enfermedad provoca cambios en el ciclo menstrual, quistes, dificultad para quedar embarazada, intensos dolores menstruales y, en el peor de los casos, cáncer de matriz [3]; sin embargo, es un tema tan poco explorado que se normaliza vivir de esa manera bajo el supuesto de que no se está enferma o que se trata de complicaciones derivadas de otros padecimientos como la hipertensión arterial o el síndrome metabólico, cuando en realidad es posible que sean consecuencias del SOP. Debido a que tiene manifestaciones clínicas muy variadas, el SOP es difícil de diagnosticar en sus primeras etapas [39], sobre todo cuando sólo se tiene acceso a servicios de salud generales.

Parte de la problemática identificada se relaciona con esquemas sociales y culturales (tabúes), dificultando aún más el acceso a los servicios de salud. Una alternativa es el uso de aplicaciones de salud de uso personal que orientan a una persona en temas de prevención y que tienen la posibilidad de sugerir la búsqueda de apoyo profesional a tiempo; sin embargo, si bien existen algunas aplicaciones específicas para la salud femenina, no se encontró ninguna que dé seguimiento al SOP, en español, y que se integre con el trabajo de médicos (as) generales o especialistas. Entre las aplicaciones disponibles enfocadas al cuidado femenino están "Modo Rosa", "Flo: Mi Calendario Menstrual", "*Aware*", entre otras [40].

Abordando el tema de las aplicaciones móviles, se tiene que, según la Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares (ENDUTIH) 2023, el 68% de la población de zonas rurales tiene acceso a telefonía móvil celular, de los cuales el 66% tiene conexión a internet. Ya que el proyecto va dirigido a la atención médica de primer nivel en el estado de Veracruz es necesario indicar que, según lo menciona el INEGI 2020, en Veracruz de Ignacio de la Llave hay 19,500 localidades rurales y 345 urbanas. A nivel nacional hay 185,243 localidades rurales y 4,189 urbanas. Dentro de dicha población 52.0% son mujeres, por tanto, en Veracruz habitan 4,190,805 mujeres.

Por otra parte, de acuerdo a la Encuesta Nacional de la Dinámica Demográfica (ENADID) 2018, independientemente de la afiliación específica a un servicio de salud, 18% de las mujeres en México se atienden en establecimientos privados (consultorio, clínica u hospital), 14% en consultorios anexos a farmacias, 32.9% en centros de salud u hospitales de la Secretaría de Salud (SSA), 27% en espacios dependientes del IMSS (Instituto Mexicano del Seguro Social) y 5.4% en el ISSSTE (Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado) [41].

1.4 Objetivo general y objetivos específicos

A continuación, se describen el objetivo general y los objetivos específicos del proyecto.

1.4.1 Objetivo general

Desarrollar una plataforma que sirva de acompañamiento a pacientes y médicos generales en clínicas de primer nivel para calcular la probabilidad de padecer SOP mediante técnicas de Inteligencia Artificial.

1.4.2 Objetivos específicos

- Analizar las aplicaciones existentes sobre la salud femenina para la identificación de las mejores prácticas.
- Investigar la sintomatología del SOP para la identificación de las variables que inciden en la probabilidad de padecer dicho síndrome.
- Generar y validar un modelo de clasificación para la probabilidad de padecer de SOP.
- Desarrollar la plataforma que permita el registro de información de pacientes para la explotación del modelo de clasificación.
- Probar la plataforma con al menos dos casos de estudio para la verificación de su funcionamiento.

1.5 Justificación

En México, la población femenina es de a 66 millones aproximadamente, de las cuales el 52% se encuentra en un rango de edad entre 15 y 49 años, este rango se denomina Edad Fértil; en ese grupo de edad, se considera que entre el 6 y el 10% de las mujeres padece SOP, es decir, entre 2 y 3 millones de personas [42]. Pese a los avances en salud, pocas mujeres tienen acceso a consultas ginecológicas para recibir el diagnóstico correcto y contar con el tratamiento a tiempo; bien sea porque los servicios públicos están rebasados o porque el servicio privado tiene un costo económico alto.

Con este trabajo se pretende ayudar en las etapas de prevención y diagnóstico a mujeres que sufren del Síndrome de Ovario Poliquístico y que no cuentan con acceso a un médico

especialista, para que tengan una mejor atención y lleven vidas más dignas en ese aspecto, al contar con una plataforma que permita: a) llevar un control de su sintomatología, b) compartirla con un médico general de primer nivel de atención (clínicas familiares, clínicas rurales) y c) hacer una predicción de diagnóstico del SOP, de tal manera que el médico decida cómo completar el diagnóstico final (realización de estudios, referir a la paciente a un especialista, iniciar un tratamiento, entre otros).

Como beneficios, se ahorrará tiempo al elaborar un diagnóstico, recursos económicos y físicos al identificar mejor los estudios clínicos y de imagen a realizar, malestares y sufrimiento a la paciente e incluso evitar procedimientos invasivos, como cirugías. Se ayudará a los(as) médicos generales en la toma de decisiones con el fin de lograr un diagnóstico más certero, ya que son más accesibles para el sector al que va orientado el proyecto. Se le dará más visualización al SOP a nivel social, ya que actualmente existen ciertos esquemas sociales y culturales que afectan el conocimiento y diagnóstico de esta enfermedad. Por parte del rubro de educación, el enfoque propuesto es aplicable a otras enfermedades, por lo que se considera adecuado el término "plataforma", con miras a incluir otras enfermedades a futuro.

En el siguiente capítulo se presenta la base teórica del presente proyecto, así como la solución planteada para el problema descrito anteriormente.

Capítulo 2 . Estado de la práctica

En este capítulo, se presentan diversos estudios, investigaciones y publicaciones relevantes que abordan aspectos clave del tema del proyecto de tesis. Esta sección proporciona una base sólida para dar contexto a la investigación propuesta y establecer su relevancia en relación con el trabajo previo realizado en el área.

2.1 Trabajos relacionados

Bharati et al. [43] mencionaron que el diagnóstico temprano del Síndrome de Ovario Poliquístico (SOP) se presentó como factor importante para evitar consecuencias graves a futuro. Por tanto, los autores propusieron un sistema capaz de ayudar en la predicción de SOP, mediante algoritmos de aprendizaje automático en un conjunto de datos específico. En concreto, aplicaron un algoritmo de selección de características univariadas para encontrar las mejores características que predijeran el SOP. De igual forma, calcularon la clasificación de los atributos y descubrieron que el atributo más importante era la proporción de hormona estimulante del folículo (FSH) y hormona luteinizante (LH). Posteriormente, aplicaron los métodos de exclusión y validación cruzada al conjunto de datos, para separar los datos de entrenamiento y de prueba. Finalmente, aplicaron a varios clasificadores, como aumento de gradiente, bosque aleatorio, regresión logística y la combinación bosque aleatorio híbrido y regresión logística (RFLR). Como resultado, demostraron que RFLR obtuvo la mejor precisión (91.01 %) y un valor de exhaustividad del 90% utilizando una validación cruzada de 40 veces aplicada a las 10 características más importantes. Por ende, RFLR fue el más adecuado para clasificar de forma fiable a las pacientes con SOP.

Actualmente, con el acceso que se tiene a internet, las personas optaron por auto diagnosticar sus enfermedades en línea en lugar de acudir a un profesional de la salud. Este hecho se convirtió en un problema conocido como *'Cibercondriasis'*. Dentro de este ámbito también se encuentran las mujeres que padecen el síndrome de ovario poliquístico. Por tanto, el estudio presentado por Soomro et al. [44] tuvo como objetivo mostrar cómo la cibercondría afectó a los pacientes que sufren de SOP. El artículo describió el estudio realizado en la región de Karachi, Pakistán, con

la participación de 100 pacientes que padecen SOP. Esto mediante análisis y regresión a los datos obtenidos de estas pacientes. Finalmente se llegó a la conclusión de que los factores de compulsión y desconfianza son determinantes para que la persona recurra a la búsqueda de información en línea sobre su enfermedad. Como conclusión, los autores mencionan que el factor de 'Compulsión' aumenta las probabilidades de búsqueda en línea.

El SOP implica la generación de quistes de varios tipos, y cada uno es diferente y representa diversos riesgos y padecimientos; es común que la presencia de quistes se diagnostique con exploraciones abdominales. Con la intención de abordar dicha problemática, Ramamoorthy et al. [45] propusieron un enfoque que determinó el síndrome de ovario poliquístico en sus primeras etapas mediante la imagen de una ecografía abdominal, esto utilizando técnicas de preprocesamiento para luego monitorear el crecimiento del quiste mediante la técnica de registro de imágenes. Con el fin de obtener el preprocesamiento de las imágenes generadas de la ecografía, los autores utilizaron enfoques existentes como Gabor, filtro gaussiano, filtro adaptativo y filtro wavelet. Sin embargo, dichos enfoques tienen limitaciones que llevaron a no obtener información completa del análisis de un quiste. Este problema se resolvió utilizando el filtro wavelet db2. Por último, el crecimiento del quiste fue monitoreado en intervalos regulares, durante los cuales se aplicaba la técnica del registro de imágenes. Dicho sistema se implementó en *MatLab*. Finalmente, el sistema ayudará a los expertos en el diagnóstico temprano del SOP con un 93% de precisión en la etapa inicial de este síndrome.

En cuanto a la clasificación, Erandathi et al. [46] previeron el índice de riesgo cronológico en personas con diabetes mellitus y sus complicaciones mediante métodos no invasivos o mínimamente invasivos, esto mediante técnicas de minería de datos y aprendizaje automático en repositorios de datos de salud. De tal manera que buscaron integrar factores faltantes en la predicción de esta enfermedad que fueron excluidos en otras investigaciones, como la combinación de datos temporales con una técnica de evaluación visual ampliamente aceptada. Por lo tanto, se propuso la introducción de una técnica de visualización simple pero eficiente para el autocontrol, que beneficiaría la gestión de la salud. Posteriormente se realizó asignando un puntaje de riesgo a cada persona basado en sus factores clínicos y biológicos, utilizando un

sistema de puntuación basado en gráficos. Finalmente, el sistema buscó impactar de manera positiva en la progresión de la diabetes y sus complicaciones previendo tempranamente el riesgo sin necesidad del conocimiento médico directo, lo que buscó ayudar a disminuir costos asociados a la diabetes mellitus.

Continuando en el ámbito de la clasificación, en el artículo de Deo et al. [47] se pretendió explorar los factores clínicos y demográficos que afectaron la incidencia de esteatosis hepática, presentando un modelo computacional que utilizó datos del NHANES-III para predecir esta condición. Con el fin de complementar el modelo, emplearon seis variables predictoras (edad, género, IMC, triglicéridos, HDL y colesterol total) y una variable de resultado (EHNA). Así mismo se abordó el desafío de datos desequilibrados mediante el algoritmo SMOTE combinado con la distancia de Gower. Posteriormente el conjunto de datos se dividió en una proporción de 70:30 para entrenamiento y prueba, respectivamente. Mediante dichos datos se entrenaron tres familias de modelos: SVM (SVM gaussiano fino y medio), *Bagged Trees* y *Boosted Tree* (*Boosted Tree* suave y ADA *Boosted Tree*). Y finalmente se utilizó la validación cruzada de 10 pliegues. Como resultado, entre los cinco modelos, el modelo "*Boosted Tree* suave" mostró una mayor precisión promedio de prueba, alcanzando un 79.03%. La sensibilidad, especificidad y AUC promedio de este modelo fueron 75.88%, 81.86% y 0.79, respectivamente.

Asimismo, Wang et al. [48] investigaron la asociación entre la historia reproductiva y el riesgo de cáncer de mama en mujeres estadounidenses. Con dicho fin, se recopilaron datos de historia reproductiva de 348 casos incidentes de cáncer de mama elegibles y 11,133 casos elegibles sin cáncer de mama entre 2009 y 2018 mediante NHANCE. Esto aplicando modelos de regresión logística incondicional y modelos de riesgos proporcionales de Cox, con los cuales se presentaron curvas de supervivencia. Como resultado identificaron una asociación negativa entre la edad en el primer parto y el riesgo de cáncer de mama, tanto en el modelo de regresión logística como en el modelo de riesgos proporcionales de Cox. Además, el número de embarazos también se encontró asociado negativamente con el riesgo de cáncer de mama en ambos modelos. Por su parte, la ooforectomía (extirpación de ambos ovarios) mostró una asociación

positiva con el riesgo de cáncer de mama en este estudio. Sin embargo, la edad en la menarquia no mostró asociación con el riesgo de cáncer de mama en este estudio.

Por otro lado, Fox et al. [49] investigaron aplicaciones móviles diseñadas para respaldar la documentación y cuantificación de datos del ciclo menstrual. El estudio propuso replantear las tecnologías relacionadas con las aplicaciones móviles, las cuales recopilaron datos íntimos sobre el ciclo menstrual, los síntomas y el comportamiento sexual de los usuarios que interactuaban con ellas, dado que buscaban predecir y gestionar el periodo menstrual. De tal manera, pretendieron ampliar el alcance del estudio en este campo mediante técnicas colaborativas de diseño. Para lograr dicho fin, se apoyaron en análisis históricos, experimentos participativos y técnicas de investigación en torno al tema del seguimiento menstrual. Finalmente, dado que es un estudio que sigue en curso, sólo identificaron la necesidad de expandir las formas de monitorear la menstruación, centrándose en la multiplicidad y dimensionalidad de la experiencia menstrual en lugar de en modelos, predictibilidad o la relación del usuario con promedios o normas.

Continuando con el ámbito de la salud femenina, el trabajo de Chopra et al. [50] tuvo como objetivo informar el diseño de tecnologías de salud inclusivas a través de la comprensión de las experiencias vividas y los desafíos de las personas con SOP. Se llevaron a cabo entrevistas semiestructuradas con 10 mujeres diagnosticadas con SOP y se analizó un foro específico de SOP en Reddit. Se informaron las prácticas de búsqueda de apoyo, construcción de sentido y auto experimentación de las personas, encontrando que la incertidumbre y el estigma fueron clave para dar forma a sus experiencias únicas con la condición. Además, se identificaron posibles vías para el diseño de tecnología que respalde sus diversas necesidades, como el seguimiento personalizado y contextual, el descubrimiento personal acelerado y la cogestión, contribuyendo así a un creciente cuerpo de literatura de HCI (*Human-Computer Interaction*) sobre temas estigmatizados en la salud y bienestar de las mujeres.

Retomando las técnicas de minería de datos, en su trabajo Wu et al. [51] mencionaron que las enfermedades cardíacas surgieron como la principal causa de muerte en los últimos tiempos,

afectando a diversas personas independientemente de su edad o género. Por tal motivo, prever el estado de salud de un paciente a tiempo es importante para iniciar un tratamiento temprano, mejorando así los posibles resultados. De tal manera que, el estudio del artículo se enfocó en cómo aprovechar las técnicas de minería de datos para desarrollar modelos de predicción de enfermedades cardíacas, esto mediante el análisis de diferentes clasificadores. Finalmente, como conclusión, los autores señalaron que la Regresión Logística y el algoritmo *Naïve Bayes* tienen una precisión más alta en conjuntos de datos de alta dimensionalidad, mientras que algoritmos como Árboles de Decisión y Bosques Aleatorios ofrecen mejores resultados en conjuntos de datos de baja dimensionalidad. También destacan que el Bosque Aleatorio proporciona una precisión superior al clasificador de Árbol de Decisión debido a que es un algoritmo de aprendizaje optimizado.

Del mismo modo, dentro de un sistema de gestión de información de estudiantes universitarios, surge el problema de que los datos se encuentren incompletos o sean de precisión ambigua. Ante este problema, Li et al. [52] propusieron minar de manera efectiva dichos datos de comportamiento de los estudiantes para mejorar su uso. Esto mediante un modelo combinado de minería de datos que utilizó algoritmos como el árbol de decisiones, la red neuronal y el algoritmo bayesiano ingenuo. Además, se estableció una plataforma de análisis y predicción del comportamiento estudiantil basada en *Spark*. Para complementar, utilizaron datos de comportamientos en el instituto, como la regularidad, hábitos de vida y condiciones de aprendizaje, como fuentes de *big data* para análisis predictivo y verificación de casos. Como resultado del estudio, mostraron que las predicciones del modelo son consistentes con la realidad, con un error promedio de predicción que no supera el 5%, validando así la efectividad del método. Finalmente, los autores señalan que los profesores tienen la posibilidad de analizar los patrones de comportamiento de los estudiantes según sus características y guiar sus conductas hacia la salud integral.

Continuando con la clasificación, se tiene que el dengue es una enfermedad amenazante causada por la picadura de un mosquito hembra. Pakistán fue víctima de esta enfermedad durante varios años, poniendo en peligro la vida de muchas personas. La detección de esta enfermedad fue un

problema debido a limitaciones de tiempo y recursos financieros. Para abordar esta situación, Jahangir et al. [53] propuso una metodología de predicción del dengue mediante técnicas de minería de datos. Esto por medio de minería de reglas de asociación, específicamente el algoritmo *Apriori* aplicado mediante la herramienta de minería de datos Weka. Posteriormente se recopiló información, incluyendo datos de pacientes con o sin dengue, para luego aplicar el algoritmo *Apriori*. Las reglas derivadas se utilizaron para generar una regla propuesta que se empleó en la predicción del dengue en pacientes. Finalmente, la regla generada pudo identificar correctamente el 75% de los registros.

En relación con el síndrome de ovario poliquístico, H. Elmannai et al. [54] mencionaron que un diagnóstico eficaz y temprano del SOP ayudó a los sistemas de atención médica a reducir las complicaciones y problemas de la enfermedad. Además, el uso de aprendizaje automático y de aprendizaje de conjunto mostró resultados prometedores en diagnósticos médicos. Por lo tanto, el objetivo principal de la investigación fue proporcionar explicaciones del modelo para garantizar eficiencia, efectividad y confianza en el modelo desarrollado a través de explicaciones locales y globales. Para esto, se utilizaron métodos de selección de características con diferentes tipos de modelos de aprendizaje automático (regresión logística, bosque aleatorio, árbol de decisiones, *Naïve Bayes*, máquina de vectores de soporte, vecinos más cercanos, *xgboost* y algoritmo *Adaboost*) con el fin de obtener una selección óptima de características y el mejor modelo. Como resultado se mostró que el apilamiento de aprendizaje automático con selección de características *Recursive Feature Elimination* (Eliminación Recursiva de Características) registró la mayor precisión en 100 en comparación con otros modelos.

Por otra parte, dentro de los criterios de diagnóstico para el síndrome de ovario poliquístico (SOP) se basaron en la opinión de expertos, el nivel más bajo de evidencia. Nunca hubo un proceso formal de consenso para determinar los criterios con el fin de llegar a un diagnóstico del SOP. S. Chang et al. [55] señalaron que los criterios de diagnóstico individuales mostraron limitaciones que resultan en una clasificación errónea de los fenotipos del SOP. Por tanto, sugirieron que es recomendable que el diagnóstico del SOP dependa de los objetivos de manejo y no necesariamente de la morfología ovárica poliquística. Enfoques de minería de datos

agnósticos identificaron subtipos de SOP que parecían genéticamente distintos, lo que potencialmente proporcionaba una forma biológicamente significativa de establecer criterios para el diagnóstico del SOP. De tal manera que, a través de análisis genéticos recientes, se sugirió que los criterios diagnósticos actuales no identificaron subtipos biológicamente distintos de SOP. Sin embargo, enfoques objetivos como el análisis de conglomerados identificaron subtipos metabólicos y reproductivos del SOP que parecieron tener arquitecturas genéticas distintas. Estos enfoques ofrecieron una transición hacia una clasificación del SOP basada en diferencias biológicas demostrables en lugar de opiniones de expertos.

También la compleja etiología del Síndrome de Ovario Poliquístico (SOP) implica, según A. E. Joham et al. [56] susceptibilidad genética, exposición a andrógenos en la vida temprana y disfunción relacionada con la adiposidad, lo que llevó a alteraciones en la función hipotalámica-ovárica. Las características clínicas del SOP son heterogéneas y llegan a manifestarse desde la adolescencia, presentándose posteriormente en la adultez a través de aspectos reproductivos, metabólicos y psicológicos. Por lo tanto, el artículo señaló que el diagnóstico del SOP utilizando los criterios de Rotterdam se basó en la presencia de dos de tres características: oligo/amenorrea, hiperandrogenismo clínico/bioquímico y morfología ovárica poliquística en ecografía, siguiendo la exclusión de causas secundarias. No obstante, las características diagnósticas del SOP variaron significativamente a lo largo de la vida reproductiva de la mujer, según el IMC y la etnia. Los criterios diagnósticos actuales dieron lugar a cuatro fenotipos distintos de SOP, y la falta de estudios longitudinales comunitarios dificultó la comprensión de su historia natural y sus características asociadas. Como resultado, los autores llegaron a la conclusión de que se necesitan más recursos para llevar a cabo estudios longitudinales bien construidos en diversas poblaciones étnicas.

Finalmente, por lo expuesto en los párrafos anteriores, S. A. Suha et al. [57] propusieron una técnica extendida de clasificación de aprendizaje automático para predecir el SOP utilizando imágenes de ecografía ovárica. Esta técnica incluyó la extracción de características de las imágenes mediante una red neuronal convolucional (CNN) y el uso de un modelo de ensamblaje de aprendizaje automático para clasificar las imágenes en criterios de SOP o no SOP. Además,

se emplearon técnicas de transferencia de aprendizaje y modelos preentrenados para mejorar la precisión y reducir la complejidad computacional. Finalmente, la técnica propuesta en el artículo demostró una precisión mejorada y una reducción en la complejidad computacional en comparación con otras técnicas existentes.

2.2 Análisis comparativo

En esta sección se presenta una comparación entre los artículos previamente consultados con el fin de identificar aspectos específicos de cada uno, los resultados se muestran mediante una tabla comparativa (Tabla 2.1).

Tabla 2.1 Análisis comparativo de los artículos relacionados

Autor	Problemática	Contribuciones	Tecnologías y/o técnicas	Resultados	Estado
Bharati et al. [43]	Identificar de manera precisa a pacientes con síndrome de ovario poliquístico (SOP) representa un desafío en el diagnóstico médico.	Identificaron las características más importantes para predecir SOP en un conjunto de datos especifico. Aplicaron algoritmos de aprendizaje automático para clasificar dichas características.	• Enfoque de diagnóstico basado en datos que involucra la aplicación de algoritmos de aprendizaje automático. • Técnicas de selección de características univariadas • Métodos de validación cruzada	El algoritmo RFLR (Regresión Logística con Bosque Aleatorio Híbrido) exhibió la mejor precisión de prueba del 91.01% y un valor de recuperación del 90% al clasificar a los pacientes con SOP.	Finalizado
Soomro et al. [44]	Comprender el impacto de la cibercondría en los pacientes con síndrome de ovario poliquístico (SOP) que buscan información de salud en línea.	Identificaron cómo la "compulsión" y la "desconfianza", son dos factores de la cibercondría, que influyen en la búsqueda de información de salud en línea por parte de pacientes con síndrome de ovario poliquístico (SOP).	• Análisis factorial exploratorio y regresión logística binaria.	El estudio encontró que el factor de 'compulsión' de la cibercondría aumentó las probabilidades de búsqueda de información de salud en línea por parte de los pacientes con SOP.	Finalizado
Ramamoorthy et al. [45]	Detectar el síndrome de ovario poliquístico (SOP) en etapas tempranas utilizando técnicas de procesamiento de imágenes para mejorar la precisión del diagnóstico.	El estudio mostro un enfoque que pudo detectar el SOP con una precisión del 93%, esto mediante imágenes de ultrasonido del abdomen, combinando estas con técnicas de registro que permitieron monitorear el crecimiento de los quistes.	• Técnicas de preprocesamiento de imágenes, como filtros wavelet. • Técnica de registro de imágenes con coeficiente de correlación y transformación afín.	Se demostró que la combinación de preprocesamiento de imágenes y técnicas de registro permitieron detectar el SOP con una precisión del 93%.	Finalizado

Autor	Problemática	Contribuciones	Tecnologías y/o técnicas	Resultados	Estado
Erandathi et al. [46]	Mejorar la predicción temprana del riesgo de diabetes mellitus y sus complicaciones a lo largo de la vida de los individuos utilizando métodos no invasivos.	El sistema buscó impactar de manera positiva en la progresión de la diabetes y sus complicaciones previendo tempranamente el riesgo sin necesidad del conocimiento médico directo, lo que buscó ayudar a disminuir costos asociados a la diabetes mellitus.	• Técnicas de minería de datos y aprendizaje automático a repositorios de datos de salud. • Sistema de puntuación de riesgo basado en gráficos.	El estudio busca predecir el porcentaje de riesgo de desarrollar diabetes, prediabetes y sus complicaciones a lo largo de la vida de un individuo, utilizando datos clínicos y biológicos para asignar un puntaje de riesgo.	En curso
Deo et al. [47]	Mejorar la capacidad predictiva de la enfermedad del hígado graso mediante técnicas de modelado y datos desequilibrados.	Desarrollo y prueba de algoritmos con datos desequilibrados para la predicción de la condición del hígado graso.	• Algoritmo SMOTE combinado con la distancia de Gower. • Técnicas de modelado computacional y aprendizaje automático.	El área bajo la curva (AUC) del modelo fue de 0.79, lo que indica un buen rendimiento en la predicción de la enfermedad del hígado graso.	Finalizado
Wang et al. [48]	Investigar la asociación entre la historia reproductiva y el riesgo de cáncer de mama en mujeres estadounidenses.	Proporciono evidencia de asociaciones entre la historia reproductiva y el riesgo de la enfermedad en mujeres estadounidenses, esto mediante métodos estadísticos avanzados.	• Modelos de regresión logística multivariable y modelos de riesgos proporcionales de Cox. • Técnicas de análisis estadístico.	Se mostró la asociación negativa entre la edad del primer parto, el número de embarazos y la edad de la menarquia entre el cáncer de mama. Por otro lado, se mostró una relación positiva entre la ooforectomía y dicha enfermedad.	Finalizado
Fox et al. [49]	Replantear la tecnología de seguimiento menstrual para promover una experiencia más inclusiva y multidimensional.	Contribuyó al reexamen y replanteamiento de la tecnología de seguimiento menstrual, explorando formas más inclusivas de interacción y conocimiento del cuerpo.	• Análisis históricos, experimentos participativos y técnicas de investigación.	Se identificó una necesidad de expandir las formas de monitorear la menstruación, centrándose en la multiplicidad y dimensionalidad de la experiencia menstrual.	En curso

Autor	Problemática	Contribuciones	Tecnologías y/o técnicas	Resultados	Estado
Chopra et al. [50]	Diseñar tecnologías inclusivas para abordar las necesidades diversas de personas con Síndrome de Ovario Poliquístico (SOP) se presenta como un desafío ante la estigmatización de la condición.	Se identificaron prácticas de búsqueda de apoyo, construcción de significado y auto experimentación, y se propusieron posibles enfoques para diseñar tecnologías que satisfagan sus diversas necesidades.	• El enfoque metodológico del estudio implica el uso de técnicas de investigación cualitativa para comprender las experiencias y desafíos de las personas con SOP.	Se identificaron oportunidades para el diseño de tecnologías de salud más inclusivas que aborden las necesidades específicas de las personas con SOP.	Finalizado
Wu et al. [51]	Predecir la supervivencia de enfermedades cardíacas es crucial para intervenir tempranamente y mejorar los resultados médicos.	Desarrollo de modelos de predicción para la supervivencia de enfermedades cardíacas mediante técnicas de minería de datos.	• Técnicas de minería de datos y aprendizaje automático.	Destacan la efectividad de diferentes clasificadores, como Regresión Logística y *Naïve Bayes*, para predecir enfermedades cardíacas en conjuntos de datos de alta dimensionalidad.	Finalizado
Li et al. [52]	Optimizar la gestión de información estudiantil en la universidad mediante la mejora de la precisión en la minería de datos.	Desarrollo de un modelo combinado de minería de datos y una plataforma de análisis y predicción del comportamiento estudiantil basada en *Spark*, utilizando algoritmos como árbol de decisión, red neuronal y algoritmo bayesiano ingenuo.	• Algoritmos de árbol de decisión, red neuronal y *Naïve Bayes* en combinación con la plataforma *Spark*.	Los resultados de predicción del modelo son consistentes con la situación real, con un error de predicción promedio que no excede el 5%.	Finalizado
Jahangir et al. [53]	Detectar el dengue en etapas tempranas sigue siendo un desafío debido a la falta de métodos eficientes y asequibles para la predicción de la enfermedad.	Presentó un método de predicción de dengue utilizando técnicas de minería de datos, específicamente el algoritmo *Apriori* de asociación de reglas	• Minería de datos • Algoritmo *Apriori* de minería de reglas de asociación	La metodología propuesta logró identificar correctamente el 75% de los casos de dengue en los registros analizados.	Finalizado

Autor	Problemática	Contribuciones	Tecnologías y/o técnicas	Resultados	Estado
H. Elmannai et al. [54]	Detectar tempranamente el síndrome de ovario poliquístico (SOP) sigue siendo un desafío para los sistemas de atención médica debido a la falta de métodos precisos y eficientes.	Presentó una metodología integral y efectiva para la detección temprana del SOP utilizando técnicas avanzadas de minería de datos y aprendizaje automático	• Técnicas de aprendizaje automático • Métodos de selección de características y técnicas de optimización bayesiana	El estudio muestra que el modelo de apilamiento de aprendizaje automático con selección de características REF logró la mayor precisión, alcanzando un 100% en comparación con otros modelos.	Finalizado
S. Chang et al. [55]	Identificar subgrupos genéticamente distintos de mujeres con síndrome de ovario poliquístico (SOP) para mejorar el diagnóstico y tratamiento.	Propuso un enfoque agnóstico de minería de datos para identificar subtipos de SOP basados en características clínicas y bioquímicas.	• Enfoques de minería de datos agnósticos. • Análisis de agrupamiento no supervisado de rasgos clínicos y bioquímicos.	Se resalta que los criterios de diagnóstico actuales para el SOP no identifican subgrupos genéticamente distintos de mujeres y que cada criterio de diagnóstico individual tiene limitaciones que llevan a la clasificación errónea entre los fenotipos del SOP.	En curso
A. E. Joham et al. [56]	Identificar y comprender la complejidad diagnóstica del síndrome de ovario poliquístico (SOP) plantea desafíos significativos en la práctica clínica.	Revisión de los criterios diagnósticos existentes y la identificación de las limitaciones en la investigación epidemiológica de largo plazo sobre el SOP.	No se aplican tecnologías o técnicas específicas, ya que se trata de una revisión exhaustiva de la literatura y de los desafíos clínicos asociados con el SOP.	Se señala la necesidad de realizar estudios longitudinales bien diseñados para mejorar el conocimiento sobre la SOP y su manejo clínico.	Finalizado
S. A. Suha et al. [57]	Automatizar la detección de múltiples quistes en ecografías ováricas para mejorar la precisión y reducir el tiempo de diagnóstico del síndrome de ovario poliquístico.	Propuso una técnica de clasificación extendida utilizando aprendizaje automático para predecir el SOP a partir de imágenes de ecografías ováricas	• Técnicas de aprendizaje profundo • Enfoque de ensamblaje de aprendizaje automático	La técnica propuesta, logró una precisión del 99.89% en la clasificación de ovarios con SOP y sin SOP utilizando imágenes de ecografías.	Finalizado

El proyecto de tesis busca complementar los estudios relacionados con el Síndrome de Ovario Poliquístico (SOP) mediante el desarrollo de una plataforma que apoye la toma de decisiones en el sector médico utilizando minería de datos y técnicas de clasificación.

Aunque en la literatura se presentan enfoques similares de clasificación del SOP mediante técnicas de Aprendizaje Automático, este proyecto está orientado hacia la atención de primer nivel en México. Esto significa que no sólo estará dirigido al especialista, como un(a) ginecólogo(a), sino que su enfoque principal será apoyar a los(as) médicos(as) generales que carecen de experiencia sobre dicha enfermedad. Así estarán en disposición de apoyar en un diagnóstico temprano a las pacientes y referirlas a tiempo a un(a) especialista, evitando retrasos en el tratamiento y gastos innecesarios en estudios.

Dirigir el trabajo hacia clínicas de primer nivel de atención es un factor importante en el desarrollo de este proyecto. En áreas rurales y semirrurales, las mujeres a menudo no tienen acceso a un(a) especialista y el SOP es una afección de difícil diagnóstico. Por tanto, este sector de la población estará en posición de obtener un diagnóstico adecuado y un tratamiento oportuno.

2.3 Propuesta de Solución

En esta sección se describe la solución a realizar en este proyecto y el proceso que la soporta, así como las tecnologías con que se implementa.

En la figura 2.1 se muestra el esquema general de la plataforma. El flujo de datos comienza con la paciente ingresando información a través de un dispositivo móvil, que se conecta mediante una API a un servidor de aplicaciones. Este servidor se comunica con una base de datos SQL, donde se almacenan datos clínicos y biomarcadores, y un proceso de minería de datos que entrena los algoritmos. Los(as) médicos, tanto generales como especialistas, acceden a una aplicación web conectada al servidor, permitiendo la entrada de datos de la paciente y facilitando la identificación de SOP.

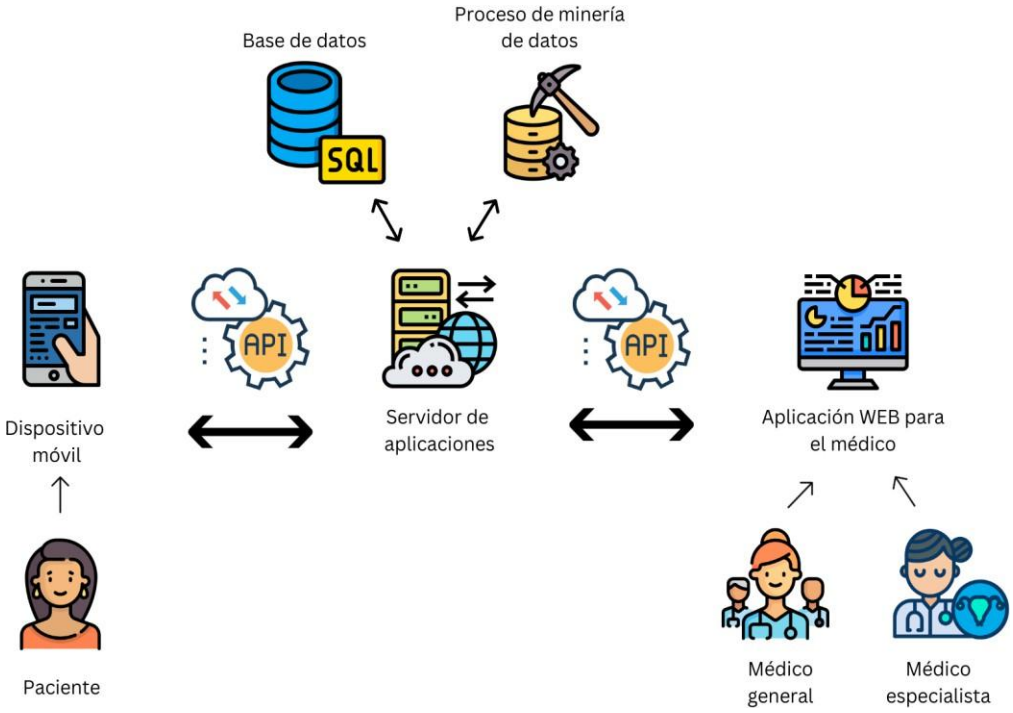

Figura 2.1 Propuesta de solución

El flujo más detallado de la plataforma, como se muestra en la figura 2.2, comienza con la paciente ingresando a su aplicación móvil, en la cual ingresará información relevante sobre su salud ginecológica, la cual comprende datos sobre su ciclo menstrual, síntomas asociados al SOP (como acné, dolor, irregularidades, etc.), antecedentes, datos generales, entre otros. Una vez ingresados los datos, el sistema ejecutara el primer modelo de minería de datos “Evaluación de necesidad de visita médica” (identificado con marcador negro), el cual se encargará de sugerirle si es necesario que acuda a una consulta con un médico general, esto considerando si hay datos anormales.

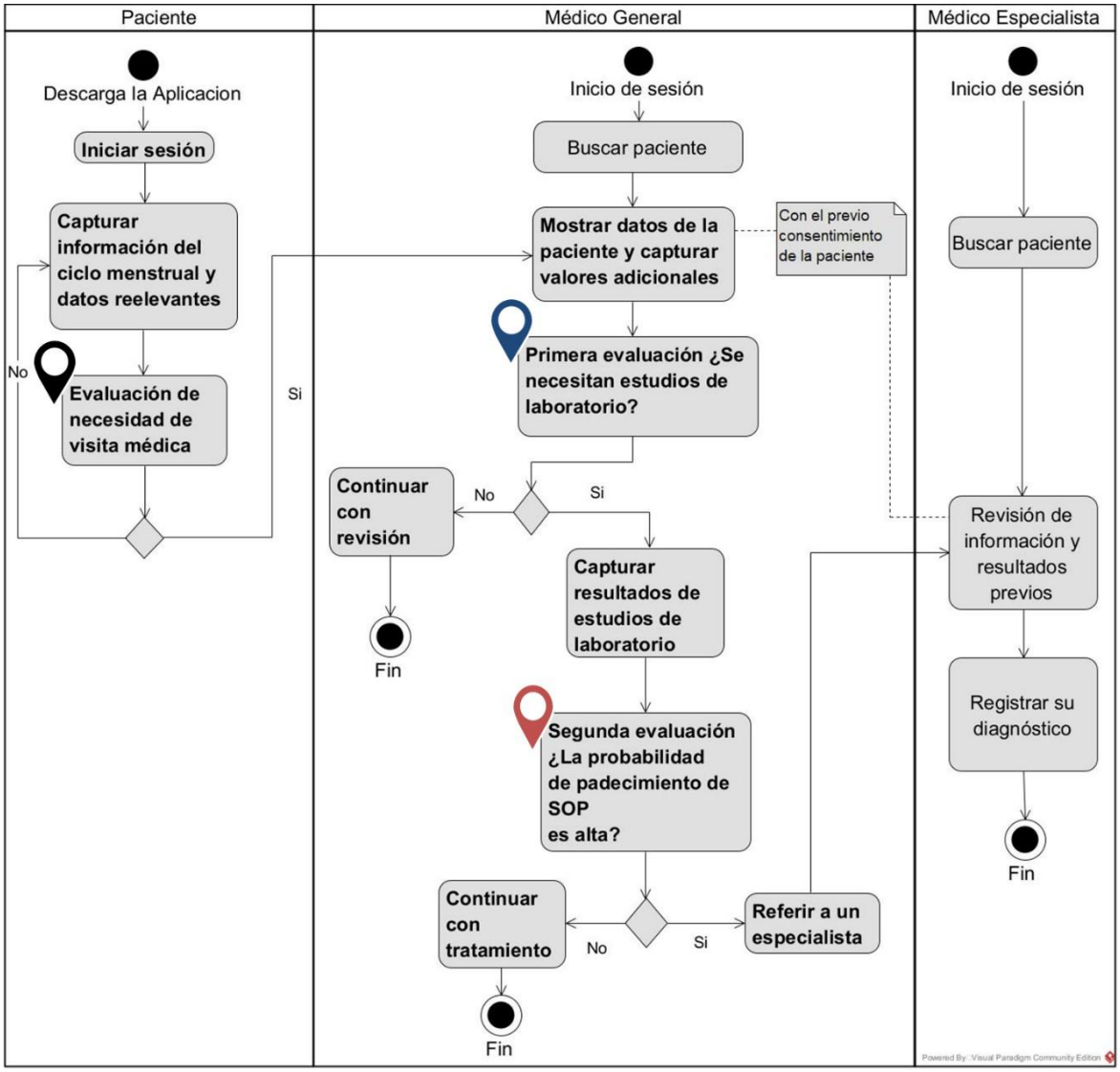

Figura 2.2 Flujo de la plataforma

Por su parte, tanto el(la) médico(a) general y el(la) especialista contarán con una aplicación web donde, previa autorización, tendrán acceso a los datos de la paciente, también contarán con la opción de ingresar más valores que sean pertinentes para complementar la información de la paciente, por ejemplo, los resultados laboratoriales y de gabinete, de manera que al final se encuentren con todas las variables necesarias con el fin de llegar a un diagnóstico adecuado. Una vez se cuente con la información de la paciente, se hará una clasificación, donde, de acuerdo a los resultados del segundo modelo "¿Se necesitan estudios de laboratorio?", se le sugerirá al

médico(a) que solicite estudios de laboratorio a la paciente (marcador azul en la figura) y, con esos valores, se hará una segunda clasificación mediante el tercer modelo "¿La probabilidad de padecer SOP es alta?" para, en su caso, sugerir la referencia a un(a) especialista (marcador rojo en la figura). El sistema pondrá a disposición de los médicos la información de la paciente, de sus estudios laboratoriales y de los resultados de clasificación previos y, a juicio del (la) especialista, se sugerirán estudios de gabinete y /o se dará el diagnóstico definitivo de SOP.

La figura 2.3 ilustra el diagrama de componentes del sistema, donde señala los elementos generales necesarios para el funcionamiento de la plataforma.

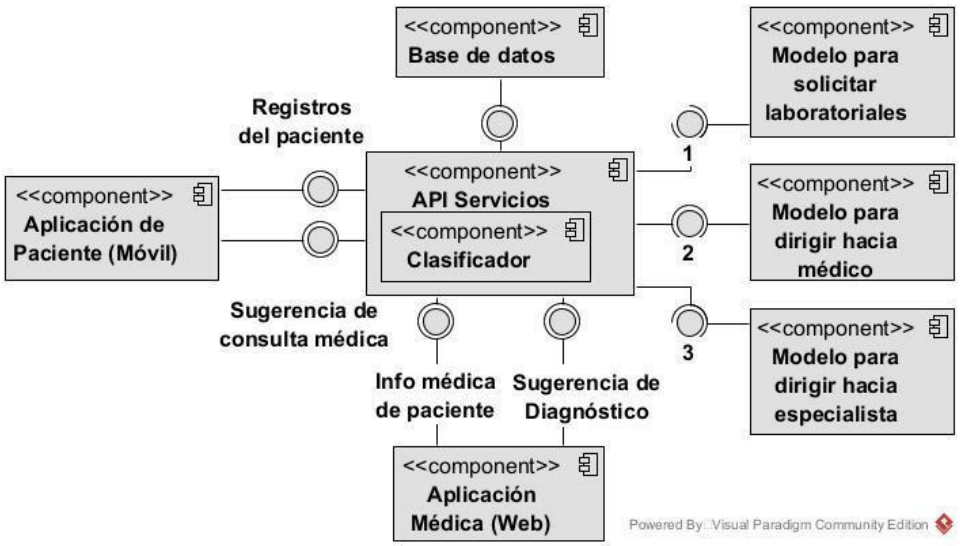

Figura 2.3 Diagrama de componentes

- Aplicación de paciente (Móvil). Este componente es una aplicación encargada de gestionar los datos proporcionados por la paciente mediante de formularios que cubren diferentes datos ginecológicos. Se comunica con el API Servicios para almacenar/consultar la información tanto de la paciente como de sus registros médicos y tiene acceso al primer modelo de minería de datos.
- API Servicios. Es el intermediario entre el almacenamiento de la información (modelos y registros de datos) y las aplicaciones tanto web como móvil. Bajo elementos de seguridad permitirá el registro y consulta de información, así como el acceso a los

modelos de Minería de Datos. Cuenta con el subcomponente Clasificador que permitirá la explotación de los modelos de clasificación mencionados anteriormente.

- Aplicación médica (Web). Mostrará los datos proporcionados por la paciente y permitirá añadir nuevos al médico general o especialista. Este componente, a través del API Servicios, hará uso de dos de los modelos de clasificación.
- Base de datos. Almacenará la información previamente gestionada por el API de servicios.

2.3.1 Justificación de la solución

En la tabla 2.2 se indican las tecnologías y herramientas que se usan en la implementación de la solución propuesta.

Tabla 2. 2 Tecnologías y Herramientas para implementar la Solución Propuesta

API Web	Marco de trabajo	Metodología	Minería	Base de Datos
Spring	React	Scrum	Weka	PostgreSQL

- Para la integración entre *back-end* y *front-end*: Spring, como marco de trabajo para el *back-end* y React, para el *front-end*, son tecnologías líderes en sus respectivos dominios. Spring ofrece un sólido soporte para el desarrollo de APIs web en Java, mientras que React proporciona una manera eficiente de construir interfaces de usuario interactivas en JavaScript, con la posibilidad de generar tanto la aplicación Web en React normal como la aplicación para dispositivo móvil en React Native, con la ventaja de que tiene la posibilidad de generar código nativo para los dos sistemas operativos más utilizados en dispositivos móviles. La integración de estas dos tecnologías asegura una comunicación fluida entre el *back-end* y el *front-end* de la aplicación.
- Metodología Ágil Probada: Scrum es una metodología ágil ampliamente adoptada en la industria del software. Su enfoque en entregas iterativas, colaboración y adaptación continua es especialmente adecuado para proyectos de desarrollo de software, permitiendo la entrega rápida de valor al cliente y la capacidad de respuesta a los cambios en los requisitos.

- Potente Análisis de Datos: Weka es una herramienta de minería de datos y aprendizaje automático que ofrece una amplia gama de algoritmos y funciones para el análisis de datos. Su versión de herramienta facilita el análisis de datos, así como el entrenamiento y validación del modelo, mientras que su versión de biblioteca para Java permite explotar fácilmente el modelo entrenado y, de ser necesario, volverlo a entrenar.
- Base de Datos Confiable: PostgreSQL es una base de datos relacional de código abierto altamente confiable y potente. Su compatibilidad con Spring y su capacidad para manejar grandes volúmenes de datos lo convierten en una opción sólida para almacenar y gestionar los datos de la aplicación de manera eficiente y segura.
- Ecosistema de Soporte y Comunidad Activa: Spring, React y PostgreSQL tienen comunidades activas y vastos ecosistemas de soporte, lo que significa que hay una gran cantidad de recursos, documentación y herramientas disponibles para ayudar en el desarrollo y mantenimiento del proyecto.

De tal manera que la solución propuesta ofrece una combinación equilibrada de tecnologías sólidas, metodología ágil probada y capacidades avanzadas de análisis de datos.

En el siguiente capítulo se describe a detalle la aplicación de la metodología para obtener la solución propuesta.

Capítulo 3 . Aplicación de la Metodología

En este capítulo se presenta cómo se aplicó la metodología de desarrollo para el sistema y el proceso seguido para generar los modelos de minería de datos a utilizar. Esta separación se debe a que se requieren tres modelos distintos para clasificar si se sugiere o no asistir al médico general, si se sugiere o no la realización de estudios de laboratorio y si se sugiere o no la derivación hacia un especialista ante el riesgo de padecer SOP; además, debido a las condiciones que debe cumplir el sistema completo (una aplicación para dispositivo móvil dirigida a pacientes y una aplicación web dirigida a profesionales de la salud), se necesita una arquitectura robusta capaz de soportar las aplicaciones y la relación con los modelos de minería de datos.

3.1 Aplicación de KDD

Como se menciona en el marco teórico, KDD permite trabajar Minería de Datos de manera formal, en este caso para obtener los tres modelos de clasificación a utilizar en la plataforma, por lo que se aplicarán los pasos tres veces.

3.1.1 Datos disponibles

De inicio, se cuenta con dos conjuntos de datos obtenidos de dos fuentes distintas:

- Datos proporcionados por la especialista en Ginecología.
  Es un conjunto de datos reducido que cuenta con 113 instancias y 6 variables, las cuales se enfoca principalmente en estudiar y clasificar la regularidad o irregularidad de los ciclos menstruales de las pacientes.
- Datos obtenidos en la plataforma *Kaggle*
  Kaggle es una plataforma de comunidad, en línea, sin costo, de ciencia de datos y que reúne científicos de datos y profesionales del aprendizaje automático, quienes registran y consumen diversos tipos de conjuntos de datos (*datasets*).
  El conjunto de datos obtenido en este medio otorga información referente al SOP, recopilada de 10 hospitales diferentes en Kerala, India, que cuenta con 542 instancias y 44 variables.

A pesar de tener esta información disponible, fue necesario recolectar datos más específicos, los cuales no están considerados dentro de las fuentes antes mencionadas, por ejemplo, los datos de mujeres sanas que no tuvieron una revisión clínica.

Se obtuvieron datos biométricos referentes a la salud de la mujer a través de encuestas realizadas a posibles usuarias de la plataforma, esto se realizó por medio de formularios de Google para aplicar cuestionarios con preguntas específicas que cubrieran las secciones necesarias para el primer modelo (permite sugerir a una mujer que visite a un médico debido a sus síntomas o, por el contrario, informarle que no parece tener problemas de salud). Dichas preguntas, al igual que la estructura del formulario, las propuso la especialista en ginecología. Posteriormente, se compartieron los cuestionarios a través de redes sociales a mujeres mexicanas que se encuentren en edad fértil, tomando en cuenta un rango de edad bastante extenso, que cubriera desde la primera menstruación de la mujer hasta la última. Las respuestas se almacenaron en un archivo CSV (variables separadas por comas).

Las preguntas se dividieron por secciones y contenían la siguiente información:

- Datos generales (8 variables)
    - Fecha de nacimiento
    - Peso
    - Altura
    - Tamaño de la cadera
    - Tamaño de la cintura
    - Sí/no consume comida rápida
    - Sí/no realiza ejercicio regularmente
    - Sí/no tiene familiares con antecedentes de hipertensión, diabetes o SOP

- Datos de indicadores de síntomas dermatológicos y hormonales (4 variables)
    - Qué nivel de oscuridad presenta en la piel
    - Sí/no presenta acné
    - Sí/no presenta crecimiento de vello corporal abundante
    - Sí/no presenta caída del cabello

- Datos ginecológicos (7 variables)
  - Qué producto menstrual utiliza
  - Cuántas veces cambia dicho producto menstrual al día
  - Si presenta dolores antes/durante/después del periodo menstrual
  - Qué cantidad de flujo vaginal presenta
  - De qué color de su flujo vaginal
  - Sí/no presenta dolor al tener relaciones sexuales
  - Método anticonceptivo que utiliza

- Antecedentes gineco-obstétricos (7 variables)
  - Edad de inicio de menstruación
  - Edad de inicio de vida sexual
  - Número de embarazos
  - Número de abortos
  - Días del ciclo menstrual
  - Días del periodo menstrual
  - Sí/no padece alguna enfermedad ginecológica diagnosticada

- Antecedentes clínicos (3 variables)
  - Presión arterial sistólica
  - Presión arterial diastólica
  - Nivel de azúcar en sangre

- Hiperandrogenismo (3 variables)
  - Crecimiento de vello corporal o hirsutismo

    Este aspecto utiliza la escala médica de Ferriman - Gallwey
    - Áreas evaluadas. La escala evalúa nueve áreas del cuerpo: labio superior, mentón, pecho, abdomen superior, abdomen inferior, pubis, brazos, muslos, y espalda.

- Puntuación: Cada área se puntúa de 0 a 4, donde 0 significa ausencia de vello y 4 representa vello denso y terminal. La puntuación total es la suma de las puntuaciones de todas las áreas evaluadas.
- Interpretación: Una puntuación total de 8 o más generalmente se considera indicativa de hirsutismo.

- Presencia de acné por áreas del cuerpo
  - Áreas evaluadas: Frente, nariz, barbilla, pecho y espalda superior
  - Puntuación:

    Grado 0: Sin lesiones.

    Grado 1: Acné leve, principalmente comedones.

    Grado 2: Acné moderado, con pápulas y algunas pústulas.

    Grado 3: Acné severo, con pústulas predominantes y algunas lesiones nodulares.

    Grado 4: Acné muy severo, con nódulos y quistes.

- Nivel de caída del cabello

  Este aspecto considera la Escala de Ludwig
  - Puntuación:

    Grado I: Adelgazamiento mínimo, especialmente en la parte superior del cuero cabelludo.

    Grado II: Adelgazamiento más pronunciado y visible.

    Grado III: Adelgazamiento severo, con una pérdida de densidad considerable en la parte superior del cuero cabelludo, a veces con el cuero cabelludo visible.

Los datos obtenidos abarcan información necesaria para los dos primeros modelos, por tanto, en el proceso KDD realizado para cada uno se definen qué variables se van a ocupar.

3.1.2 Proceso KDD para el primer modelo

El primer modelo se utiliza en la aplicación para dispositivo móvil de la paciente y analiza los datos previamente ingresados por ella (datos generales, antecedentes gineco-obstétricos y heredofamiliares), esto con el fin de que se le sugiera si es necesario que acuda con un médico general. Para este modelo se considera que el sistema no evalúe si es posible que la paciente presente alguna enfermedad ginecológica particular que requiera atención médica, como SOP, endometriosis, síndrome premenstrual, dismenorrea, entre otras. De manera que sólo se sugerirá a la paciente que acuda con un médico, ya que sus datos presentan anomalías, sin indicar directamente qué enfermedad es posible que presente, ya que este modelo sólo señala si tiene un problema o no.

3.1.2.1 Limpieza de datos

La información que se obtuvo de los cuestionarios digitales se procesó mediante técnicas de limpieza y transformación, lo cual permitió eliminar inconsistencias, sustituir valores faltantes y adaptar las variables. Dicho procesamiento fue fundamental para mejorar la calidad de los datos y obtener resultados más precisos durante el análisis.

El proceso se encontró dividido en las siguientes partes:

- Selección de variables y su justificación

  En este punto se seleccionan las variables necesarias, ya que cada modelo tiene requerimientos distintos, por tanto, se separa la información de acuerdo a las necesidades. Dentro de las encuestas se agregaron puntos más específicos que son necesarios para el segundo modelo, sin embargo, al no ser requerida esa información para el primero es necesario hacer una separación. Para este modelo se excluyen las variables pertenecientes a estas secciones de la encuesta:

  - Antecedentes clínicos
  - Hiperandrogenismo

  Todas las demás variables consideradas en la encuesta son incluidas en este modelo (33 variables), que pertenecen a las secciones:

  - Datos generales

- Datos de Indicadores de Síntomas Dermatológicos y Hormonales
- Datos ginecológicos
- Antecedentes gineco-obstétricos

- Procesamiento y transformación de datos

  El preprocesamiento consta de encontrar y modificar los valores inconsistentes o nulos que se identifiquen dentro del conjunto de datos, ya que es posible que existan errores de captura al momento de ingresar la información.

  Por otra parte, la transformación hace referencia al proceso de adaptar los datos con el fin de que sean más manejables, estas transformaciones dependen tanto del tipo de dato específico como del ámbito de trabajo particular; por ejemplo, la edad de inicio de menstruación es un número entero con un dominio del 7 al 20, sin embargo para cuestiones de salud, es suficiente con saber si el valor es menor a 9, entre 9 y 15 o mayor a 15, de manera que, en el caso particular de la edad de inicio de menstruación, los valores se discretizan.

  Dentro de las variables modificadas se tienen las siguientes:

Tabla 3.1 Preprocesamiento y transformación de datos primer modelo

Variable	Preprocesamiento	Transformación
Fecha de nacimiento (formato dd/MM/aaaa)	Si es posible, se interpreta la fecha. De lo contrario, se elimina el registro.	Se calcula la edad numérica. Se discretiza (menor a 18, 18-24, 25-30, 31-35, mayores de 35).
Peso en kilogramos	Si se encuentra fuera de rango (<30 kilogramos, mayor a 500 kilogramos) se elimina el registro.	No requiere.
Altura en centímetros	Si se encuentra fuera de rango (menor a 100 cm, mayor a 200 cm) eliminar el registro.	Se ajusta en los casos donde se captura en metros

Variable	Preprocesamiento	Transformación
IMC	Se realiza el cálculo directo en el archivo CSV, donde se consigue a partir del peso y altura (peso en kilogramos dividido por el cuadrado de la estatura en metros)	Se discretiza (Bajo peso: <18.5, Peso normal: 18.5-24.9, Sobrepeso: 25-29.9, Obesidad grado I: 30-34.5, Obesidad grado II: 35-39.9, Obesidad grado III: >40).
Tamaño de cadera en centímetros	Si se encuentra fuera de rango (<60 cm, mayor a 150 cm), eliminar el registro.	No requiere.
Tamaño de cintura en centímetros	Si se encuentra fuera de rango (<40 cm, mayor a 150 cm), eliminar el registro.	No requiere.
Proporción cintura/cadera	Se realiza el cálculo directo en el archivo CSV, donde se consigue a partir del tamaño de la cintura y tamaño de la cadera.	No requiere.
Caída del cabello	No requiere ya que son valores con respuestas delimitadas (seleccionables) y no abiertas.	Se colocó una clave para cada valor (No presenta caída del cabello: 0, Sí presenta caída del cabello: 1 poca caída de cabello, 2 caída de cabello considerable, 3 caída de cabello abundante).
Problemas de acné	No requiere ya que son valores con respuestas delimitadas (seleccionables) y no abiertas.	Se colocó una clave para cada valor (No presenta acné: 0, Si: 1 presenta acné leve, 2 presenta acné moderado, 3 presenta acné grave, 4 presenta acné muy grave).

Variable	Preprocesamiento	Transformación
Producto menstrual	Combinación de producto de higiene menstrual que utiliza la paciente y cantidad de veces que se cambia dicho producto al día durante su periodo menstrual.	Discretizar a tres valores: poco abundante, normal y muy abundante, de acuerdo a las indicaciones de la ginecóloga. Por ejemplo: Uso de toallas sanitarias, si es nocturna o normal. Cambio de menos de 3 veces al día indica normal, de más de 4
	cambia dicho producto al día durante su periodo menstrual.	veces al día indica abúndate, más de 5 veces al día indica que es muy abundante.
Método anticonceptivo	No requiere ya que son valores con respuestas delimitadas (seleccionables) y no abiertas.	No requiere.

Las variables que no fueron incluidas dentro de la tabla no requieren preprocesamiento ni transformación, ya que, dentro de la encuesta, esos datos tenían respuestas delimitadas seleccionables.

3.1.2.2 Selección del algoritmo de minería

Después de seleccionar, procesar y transformar la información, se procedió a entrenar el primer modelo, utilizando la herramienta Weka y aplicando los siguientes algoritmos de clasificación: *Naïve Bayes* [14], Perceptrón Multicapas (*Multilayer Perceptron*) [15], IBK [16], J48 [17] y *Random Forest* [18]; se aplicaron estos algoritmos porque, en diversos artículos consultados para el desarrollo del estado del arte, se encuentran entre los más utilizados y con mejores resultados [57].

Para evaluar la precisión de los algoritmos para los datos del primer modelo, se tomaron en cuenta los siguientes parámetros: porcentaje de instancias clasificadas correctamente, F1 y la matriz de confusión generada (de acuerdo a lo analizado en la sección 1.1.2 de este documento).

El conjunto de datos obtenido a partir de las encuestas consta de 103 instancias y las 33 variables mencionadas anteriormente; del total de instancias, 69 corresponden a mujeres sanas y 34 a mujeres con alguna enfermedad ginecológica diagnosticada. El atributo objetivo que se desea clasificar es "Enfermedad Ginecológica" (Sí tiene / No tiene). En la tabla 3.2 se muestran los resultados obtenidos por cada algoritmo.

Tabla 3.2 Resultados por algoritmo para el primer modelo

	Naïve Bayes	**Multilayer Perceptron**	**IBK**	**J48**	**Random Forest**
Clasificadas correctamente	75%	74.0385%	75.9615%	78.8462%	76.699%
Clasificadas incorrectamente	25%	25.9615%	24.0385%	21.1538%	23.301%

A pesar de que el porcentaje de instancias clasificadas correctamente está en un nivel aceptable (mayor al 70%), las clases son desbalanceadas debido a que lo más común es que las mujeres sean sanas; por lo tanto, la cantidad de instancias clasificadas correctamente no es una medida suficiente para elegir el mejor algoritmo; por esa razón se calcularon también los valores de precisión, sensibilidad y especificidad que se explicaron en el Marco Teórico.

Como ejemplo ilustrativo se presentan los cálculos para el algoritmo de Árboles de Decisión (J48) que presentó el mejor porcentaje de aciertos:

- Precisión o exactitud: 82 instancias fueron clasificadas correctamente.
- Sensibilidad: 8 instancias fueron clasificadas correctamente como positivas a tener una enfermedad ginecológica.

- Especificidad: 74 instancias fueron clasificadas correctamente como negativas a tener una enfermedad ginecológica.

En la tabla 3.3 se muestran los resultados del cálculo de cada medida para el algoritmo J48 en el primero modelo.

Tabla 3.3 Resultados obtenidos por el algoritmo J48 para el primer modelo

Datos	Exactitud	Sensibilidad	Especificidad
104 registros	78.85%	38.1%	89.2%

En la matriz de confusión del algoritmo J48 (Tabla 3.4) se muestra la diferencia significativa en la cantidad de valores entre clases.

Tabla 3.4 Matriz de confusión de algoritmo J48 con 104 registros para el primer modelo

Matriz de confusión de J48		
No está enferma	**Sí está enferma**	
74 registros	9 registros	**No está enferma**
13 registros	8 registros	**Sí está enferma**

Teniendo todos estos factores en cuenta, se identificó la necesidad de mejorar los resultados mediante distintas técnicas. Como primera alternativa, se revisaron las variables para identificar si algunas eran más relevantes para el modelo. En el primer entrenamiento se consideraron 33 variables, posteriormente se seleccionaron las 6 más influyentes mediante el algoritmo evaluador `CfsSubsetEval`, utilizando el método *Best First*, aunque los resultados mejoraron, no fue suficiente debido a que las clases se encuentran desbalanceadas. Por lo tanto, a la nueva selección se le aplicó el algoritmo SMOTE-NC para balancear los datos, tanto categóricos como continuos. Este último proceso agregó 34 instancias sintéticas, dando un total de 138 registros y 5 variables.

Con el nuevo conjunto de datos, se volvió a entrenar el modelo, en la tabla 3.5 se muestran los resultados obtenidos por el algoritmo J48. Se observa que, una vez balanceadas las clases, la exactitud mejoró respecto a los resultados anteriores.

Tabla 3.5 Resultados obtenidos por el algoritmo J48 con 138 registros para el primer modelo

Datos	Exactitud	Sensibilidad	Especificidad
138 registros	81.88%	87.0%	76.8%

A partir de estos resultados se sometió el conjunto de datos nuevamente a diversas selecciones de variables mediante métodos de filtrado como el mencionado anteriormente, esto para tener mejor certeza de cuáles variables son las más relevantes para lograr obtener un modelo más consistente.

Dado que los datos propuestos son principalmente categóricos, la selección de características se enfocó en técnicas que funcionaran bien con atributos discretos y permitieran mejorar el rendimiento del modelo. Para ello, se probaron cuatro métodos de filtrado que se adaptaron bien a datos categóricos, de esta selección se obtuvieron las variables indicadas en la tabla 3.6:

Tabla 3.6 Selección de atributos primer modelo

Método	Atributos más relevantes (Top 6)
Chi-Square	DolorSexo, OscPiel, Durante, GrupoEdad, SOP, GrupoIMC
InfoGain	DolorSexo, OscPiel, Durante, GrupoEdad, SOP, EdadIniMenst
CfsSubsetEval	OscPiel, Durante, DolorSexo, Embarazos, EdadIniMenst, SOP
ReliefF	DolorSexo, Antes, OscPiel, Ejercicio, Durante, Condón

Se identifican con todos los métodos
Se identificas con 3 métodos
Se identifican con 2 métodos

Después de ejecutar los algoritmos de selección de variables, de acuerdo a la tabla 3.6, se eligió utilizar sólo las que fueron seleccionadas por los cuatro algoritmos.

- Conjunto de datos con 138 registros y 6 variables
    - Dolor durante relaciones sexuales
    - Oscuridad en piel

- Dolor durante la menstruación.
- Grupo de edad
- SOP en familiares
- Edad de inicio de menstruación.

Se volvieron a ejecutar los cuatro algoritmos de clasificación anteriormente mencionados, pero ahora con las nuevas variables seleccionadas y las clases balanceadas, mostrando así una mejoría en el resultado como se ve en las figuras 3.1 y 3.2.

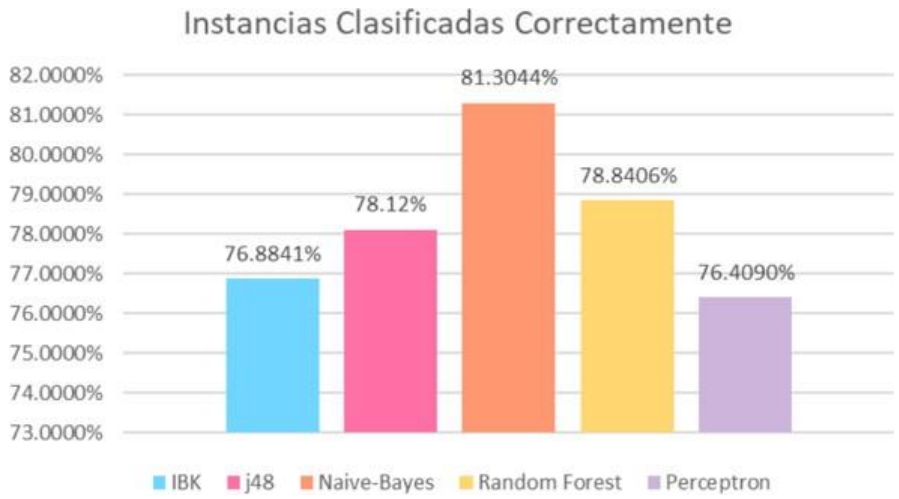

Figura 3.1 Gráfica de comparación entre algoritmos clasificadores para instancias clasificadas correctamente en el primer modelo

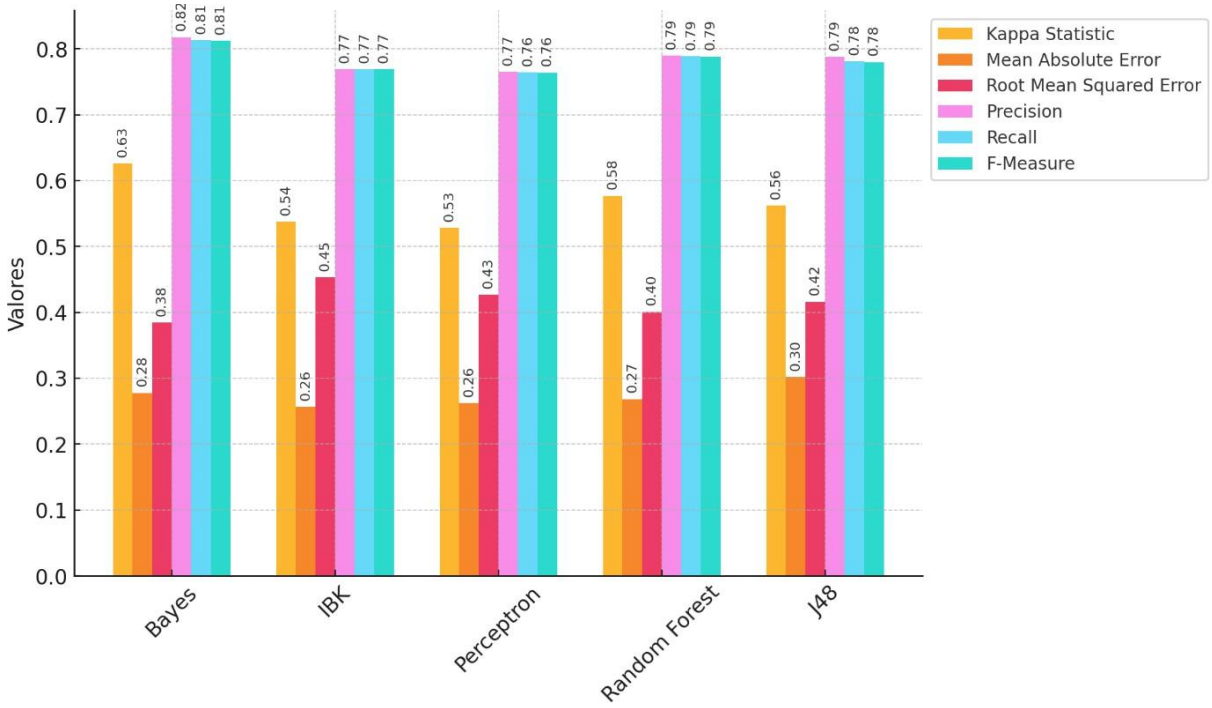

Figura 3.2 Gráfica de comparación entre algoritmos clasificadores para parámetros de error y precisión en el primer modelo

Por lo anterior, se considera que el algoritmo *Naïve Bayes* obtuvo mejores resultados y el modelo de clasificación resultante es el que se utiliza en la plataforma para sugerir a la paciente que acuda a un médico general al presentar un riesgo de padecer alguna enfermedad ginecológica. Es importante hacer notar que, para cada algoritmo, se realizaron diez entrenamientos con diez semillas distintas, esto mediante *Cross-validation* con 10 separaciones (*folds*). Se registraron los parámetros a evaluar y de cada uno de ellos se obtuvieron la media y la desviación estándar, ya que los especialistas en minería de datos indicaron durante las estancias profesionales que dichas medidas permiten describir de manera clara el comportamiento central y la variabilidad de los resultados obtenidos por el modelo; estas medidas fueron suficientes debido a que los datos no presentaron valores atípicos importantes ni distribuciones sesgadas que justificaran otro enfoque. A continuación, se muestran los valores para el algoritmo *Naïve Bayes*, el cual mostró los mejores resultados (gráficas anteriores).

Tabla 3.7 Resultados primer modelo en algoritmo Naïve Bayes

Cross-validation							
Semilla	**Instancias clasificadas correcta-mente (%)**	**Estadís-tica Kappa**	**Error absoluto medio**	**Raíz del error cuadrá-tico medio**	**Preci-sión**	**Sensibi-lidad**	**Medida F**
1	82.6087%	0.6522	0.2764	0.3817	0.831	0.826	0.826
2	81.1594%	0.6232	0.2785	0.3858	0.814	0.812	0.811
3	83.3333%	0.6667	0.2759	0.384	0.837	0.833	0.833
4	78.2609%	0.5652	0.2826	0.3877	0.786	0.783	0.782
5	78.9855%	0.5797	0.2807	0.3871	0.793	0.790	0.789
6	82.6087%	0.6522	0.2713	0.3777	0.831	0.826	0.826
7	81.8841%	0.6377	0.279	0.385	0.824	0.819	0.818
8	81.8841%	0.6377	0.2788	0.3843	0.824	0.819	0.818
9	81.1594%	0.6232	0.2808	0.3883	0.818	0.812	0.811
10	81.1594%	0.6232	0.2754	0.3842	0.816	0.812	0.811
Media	81.3044%	0.6261	0.27794	0.38458	0.8174	0.8132	0.8125
Desvia-ción estándar	1.59%	0.031914	0.003278	0.003122	0.016453	0.015781	0.016105

3.1.3 Proceso KDD para el segundo modelo

Para el caso del segundo modelo, donde se requiere que la aplicación web analice los datos obtenidos de la aplicación móvil de la paciente y los datos ingresados en la aplicación web por el médico (datos clínicos e hiperandrogenismo), con el fin de sugerir al médico si es necesario realizar estudios de laboratorio, debido a que es probable que padezca SOP y se necesiten más datos para confirmar un diagnóstico adecuado. De manera que sólo se sugerirá al médico que solicite estudios laboratoriales si la probabilidad de padecer SOP es alta.

3.1.3.1 Limpieza de encuestas

La información necesaria para el segundo modelo fue adquirida en la segunda encuesta realizada por redes sociales, la cual de igual forma que con la primera encuesta, contiene preguntas propuestas por la especialista en ginecología, con la diferencia de que en este caso el formulario buscó centrarse más en datos clínicos, así como reducir las respuestas abiertas y priorizar las opciones múltiples delimitando mejor los resultados para obtener datos más limpios. De tal manera que el proceso de limpieza se realizó de la siguiente forma:

- Selección de variables y su justificación

  El segundo modelo analiza los datos obtenidos tanto por la paciente como los proporcionados por el médico general, por lo que las variables necesarias para este modelo se encuentran en las siguientes secciones:

  - Datos generales
  - Datos de Indicadores de Síntomas Dermatológicos y Hormonales
  - Datos ginecológicos
  - Antecedentes gineco-obstétricos
  - Antecedentes clínicos
  - Hiperandrogenismo

  Sin embargo, dentro de estas secciones se encontraban las variables referentes a métodos anticonceptivos, las cuales causaban ruido ya que no eran relevantes para el modelo, de tal forma que la especialista en ginecología sugirió discretizarlas para obtener mejores resultados.

  - Condón
  - DIU
  - Inyección
  - Parche
  - Pastilla
  - Operación
  - Implante

- Procesamiento y transformación de datos

  Para este modelo también hubo necesidad de hacer tanto un preprocesamiento como transformación de datos. Debido a que varios de los atributos también existen en el primer modelo, aquí sólo se describen los ajustes realizados en los atributos exclusivos del segundo modelo (tabla 3.8).

Tabla 3.8 Preprocesamiento y transformación de datos segundo modelo

Variable	Preprocesamiento	Transformación
Caída del cabello por área del cuerpo	No requiere ya que son valores con respuestas delimitadas (seleccionables) y no abiertas.	No requiere.
Hirsutismo por área del cuerpo	No requiere ya que son valores con respuestas delimitadas (seleccionables) y no abiertas.	Se discretiza la suma de los valores por área del cuerpo (<6 poco probable SOP, 6-8 posible SOP, pero requiere de otros síntomas, >8 SOP muy probable)
Problemas de acné por área del cuerpo	No requiere ya que son valores con respuestas delimitadas (seleccionables) y no abiertas.	Se discretiza la suma de los valores por área del cuerpo (<6 poco probable SOP, 6-9 posible SOP, pero requiere de otros síntomas, >10 SOP muy probable)
Edad de inicio de relaciones sexuales	No requiere ya que son valores con respuestas delimitadas (seleccionables) y no abiertas.	Se discretiza (No tiene relaciones: 0, Menor a 18 años: 1, Entre 19 y 25 años: 2, Mayor a 25: 3).

Las variables que no fueron incluidas dentro de la tabla y que no se consideraron en el primer modelo no requieren preprocesamiento ni transformación, ya que, dentro de la encuesta, esos datos tenían respuestas delimitadas y no se prestan a errores de captura.

3.1.3.2 Selección del algoritmo de minería

Una vez que las variables se procesaron, se realizó el entrenamiento del modelo dentro de la herramienta Weka, para este modelo se consideraron los mismos algoritmos de clasificación y medidas que en el primer modelo.

El conjunto de datos obtenido a partir de las encuestas consta de 56 registros y 26 variables, entre las cuales se encuentran las consideradas para el primer modelo, solo se omiten aquellas que son referentes a métodos anticonceptivos, dado que en este modelo no son necesarios. El atributo objetivo que se desea predecir es "Necesidad de datos laboratoriales" (Sí es necesario / No es necesario).

En la tabla 3.9 se muestran los resultados de los algoritmos para el conjunto de datos del modelo que clasifica la necesidad de solicitar estudios laboratoriales.

Tabla 3.9 Resultados de algoritmos para segundo modelo

	Naïve Bayes	***Multilayer Perceptron***	**IBK**	**J48**	***Random forest***
Clasificadas correctamente	75%	64.2857%	53.5714%	64.2857%	66.0714%
Clasificadas incorrectamente	25%	35.7143%	46.4286%	35.7143%	33.9286%

Nuevamente, el conjunto de datos presenta clases desbalanceadas, hay más mujeres sanas o con una enfermedad distinta a SOP que, por lo tanto, no requieren estudios laboratoriales, por lo que la cantidad de instancias clasificadas correctamente no es suficiente para elegir un algoritmo.

Como el algoritmo *Naïve Bayes* presenta el mejor porcentaje de instancias clasificadas correctamente, se calcularon las otras medidas para sus resultados:

- Precisión o exactitud: 42 registros fueron clasificadas correctamente de un total de 56.
- Sensibilidad: 7 registros fueron identificados correctamente como pacientes con posible SOP y necesidad de estudios laboratoriales.
- Especificidad: 35 registros fueron correctamente reconocidos como posibles negativos a SOP y por tanto sin necesidad de estudios laboratoriales.

En la tabla 3.10 se muestra el resultado de las medidas calculadas para el algoritmo *Naïve Bayes* en el segundo modelo.

Tabla 3.10 Resultados obtenidos por el algoritmo *Naïve Bayes* con 56 registros para el segundo modelo

Datos	Exactitud	Sensibilidad	Especificidad
56 registros	75%	41.2%	89.7%

En la matriz de confusión del algoritmo *Naïve Bayes* (Tabla 3.11) se muestra la diferencia significativa en la cantidad de valores entre clases.

Tabla 3.11 Matriz de confusión de algoritmo *Naïve Bayes* con 56 registros para el segundo modelo

Matriz de confusión de *Naïve Bayes*		
No SOP	**Posible SOP**	
35 registros	4 registros	**No SOP**
10 registros	7 registros	**Posible SOP**

Teniendo estos factores en cuenta, al igual que en el primer modelo, es necesario ajustar los datos. Por tanto, se aplicó el algoritmo SMOTE-NC, para balancear los datos tanto categóricos como continuos, lo cual agregó 22 registros sintéticos, dando un total de 78 registros y conservando las 26 variables iniciales.

Se realizó el entrenamiento de la misma manera que con el primer modelo y se obtuvo que el algoritmo que presentó mejores resultados fue *Random Forest* (figuras 3.3 y 3.4), los valores

obtenidos de 10 semillas independientes para *Cross-validation* con 10 separaciones arrojaron los resultados que se muestran en la tabla 3.12:

Tabla 3.12 Resultados segundo modelo en algoritmo *Random Forest*

	Cross-validation						
Semilla	**Instancias clasificadas correctamente (%)**	**Estadística Kappa**	**Error absoluto medio**	**Raíz del error cuadrático medio**	**Precisión**	**Sensibilidad**	**Medida F**
1	87.1795%	0.7436	0.2839	0.3442	0.876	0.872	0.871
2	84.6154%	0.6923	0.2728	0.3348	0.847	0.846	0.846
3	88.4615%	0.7692	0.2731	0.3297	0.885	0.885	0.885
4	91.0256%	0.8205	0.2751	0.3295	0.913	0.910	0.910
5	84.6154%	0.6923	0.2772	0.341	0.847	0.846	0.846
6	88.4615%	0.7692	0.2725	0.3338	0.885	0.885	0.885
7	88.4615%	0.7692	0.2714	0.3286	0.887	0.885	0.884
8	87.1795%	0.7436	0.2851	0.3412	0.876	0.872	0.871
9	88.4615%	0.7692	0.271	0.3322	0.887	0.885	0.884
10	88.4615%	0.7692	0.2668	0.3221	0.885	0.885	0.885
Media	87.6923%	0.75383	0.27489	0.33371	0.8788	0.8771	0.8767
Desviación estándar	1.93%	0.038596	0.005743	0.006810	0.0196	0.019382	0.019356

Para mostrar los resultados obtenidos en los demás algoritmos clasificadores se generaron las figuras 3.3 y 3.4, las cuales que ilustran el comportamiento en cada parámetro considerado.

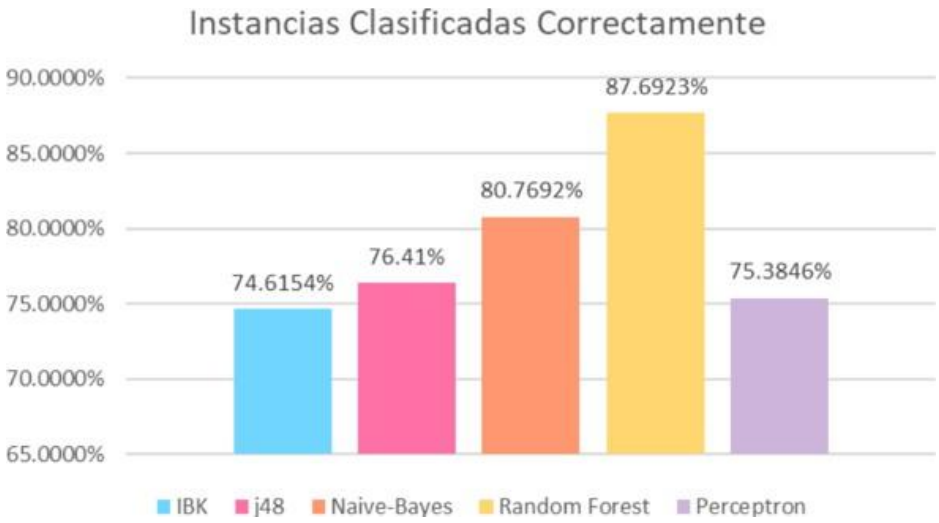

Figura 3.3 Gráfica de comparación entre algoritmos clasificadores para instancias clasificadas correctamente en el segundo modelo

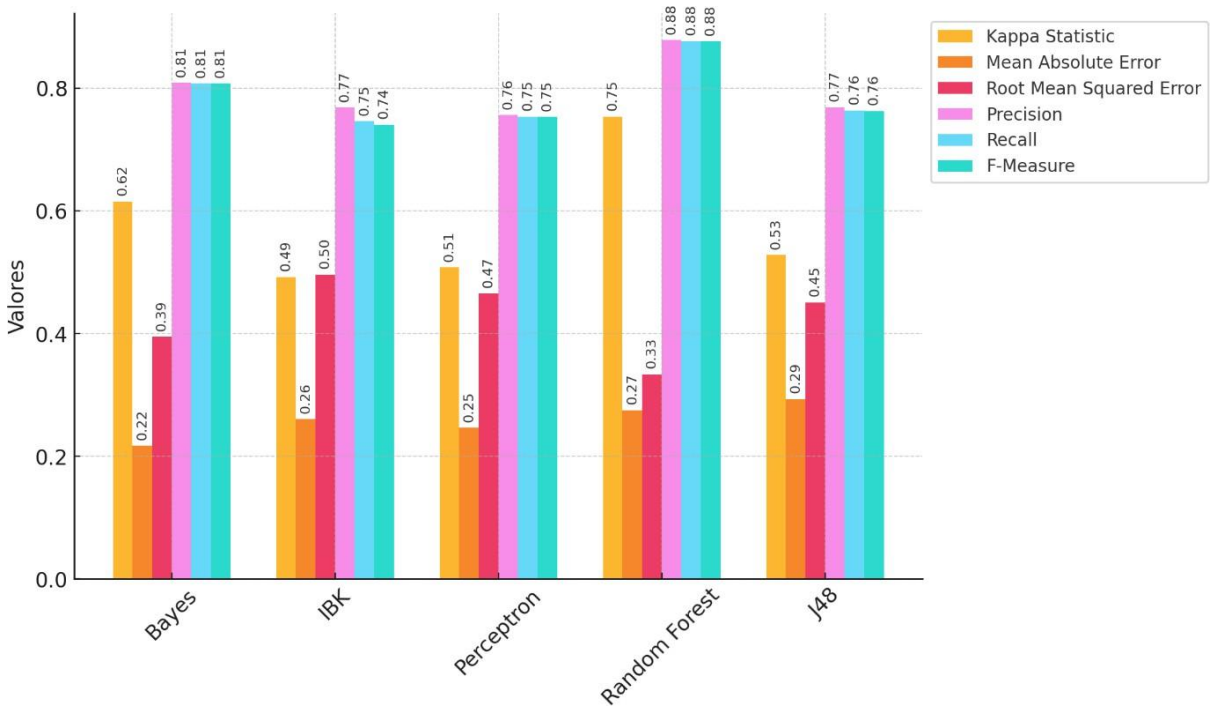

Figura 3.4 Gráfica de comparación entre algoritmos clasificadores para parámetros de error y precisión en el segundo modelo

3.1.4 Proceso KDD para el tercer modelo

El tercer modelo requiere que la aplicación web analice los datos proporcionados por la paciente, los datos clínicos y los datos laboratoriales ingresados por el médico, de manera que se realice una evaluación general de los mismos y que se indique qué tan probable es que la paciente padezca SOP. En caso de que la probabilidad que se encuentre sea alta, el sistema sugerirá al médico general derivar a la paciente con un médico especialista con el fin de que reciba un diagnóstico definitivo.

3.1.4.1 Obtención de datos

Los datos necesarios para este modelo se obtuvieron a través de la plataforma *Kaggle*, la cual contiene datos ginecológicos, obstétricos, generales y laboratoriales, que abarcan parte de la información necesaria para la evaluación de este tercer modelo; el resto de la información provino de las encuestas utilizadas en los modelos anteriores.

- Selección de variables y su justificación

  Ya que se trata de un conjunto de datos (*dataset*) probado anteriormente, no requirió de mayores cambios, sin embargo, se eliminaron variables que por la naturaleza del proyecto no estaban contempladas. Las variables seleccionadas para este modelo y que son parte del conjunto de datos son las siguientes:

  - Edad
  - Peso
  - Altura
  - IMC (Índice de Masa Corporal)
  - Duración del ciclo
  - Embarazos
  - Abortos
  - FSH (Hormona foliculoestimulante)
  - LH (Hormona luteinizante)
  - FSH/LH (Relación Hormona foliculoestimulante / Hormona luteinizante)

- Longitud de la Cadera
- Longitud de la Cintura
- Relación Cintura/cadera
- TSH (hormona estimulante de la tiroides)
- Prolactina
- Progesterona
- Azúcar en sangre
- Crecimiento de vello
- Oscuridad en piel
- Caída de cabello
- Acné
- Consume comida rápida o no
- Realiza ejercicio o no
- Presión arterial sistólica
- Presión arterial diastólica

- Procesamiento y transformación de datos

Las variables que son propias de este modelo, como lo son los resultados hormonales y los niveles de azúcar en sangre, no requirieron ni preprocesamiento ni transformación; para otros datos como la edad o el peso, ya se describió su tratamiento en los modelos anteriores.

3.1.4.2 Selección del algoritmo de minería de datos

Una vez que se realizaron los ajustes al conjunto de datos se continuó con el entrenamiento del tercer modelo. El conjunto de datos obtenido de la plataforma *Kaggle* consta de 541 registros y 22 variables, entre las que se encuentran datos generales, gineco-obstétricos, hormonales y clínicos. El atributo objetivo que se desea clasificar es “Probabilidad de SOP” (Sí es probable / No es probable).

En la tabla 3.13 se muestran los resultados de los algoritmos para el conjunto de datos del modelo para identificar SOP.

Tabla 3.13 Distribución de datos por algoritmo tercer modelo

	Naïve Bayes	***Multilayer Perceptron***	**IBK**	**J48**	***Random forest***
Clasificadas correctamente	59.7043%	78.7431 %	78.5582 %	80.2218 %	83.549 %
Clasificadas incorrectamente	40.2957 %	21.2569 %	21.4418 %	19.7782 %	16.451 %

A pesar de que el porcentaje de instancias clasificadas correctamente está en un buen nivel con excepción del caso del algoritmo *Naïve Bayes*, al tener una menor cantidad de registros de mujeres enfermas con SOP respecto a las que no lo están, la medida de instancias clasificadas correctamente no es suficiente para elegir un algoritmo.

Teniendo estos factores en cuenta, es necesario ajustar los datos para mejorar el comportamiento del modelo. Por tanto, se aplicó el algoritmo SMOTE para balancear los datos, lo que agregó 177 datos sintéticos, dando un total de 718 registros y conservando las 22 variables iniciales, ya que la selección de variables no representaba un cambio en el desempeño del modelo.
Se realizó el entrenamiento de la misma manera que con los modelos anteriores, obteniendo que el algoritmo que presentó mejores resultados fue *Random Forest* (figuras 3.8 y 3.9), los valores obtenidos de 10 semillas independientes para *Cross-validation* con 10 separaciones arrojaron los resultados que se muestran en la tabla 3.14

Tabla 3.14 Resultados tercer modelo en algoritmo *Random Forest*

	Cross-validation						
Semilla	**Instancias clasificadas correctamente (%)**	**Estadística Kappa**	**Error absoluto medio**	**Raíz del error cuadrático medio**	**Precisión**	**Sensibilidad**	**Medida F**
1	84.8189%	0.6962	0.2772	0.3434	0.848	0.848	0.848
2	84.8189%	0.6962	0.2754	0.3409	0.848	0.848	0.848

Cross-validation							
Semilla	**Instancias clasificadas correctamente (%)**	**Estadística Kappa**	**Error absoluto medio**	**Raíz del error cuadrático medio**	**Precisión**	**Sensibilidad**	**Medida F**
3	83.9833%	0.6795	0.2764	0.3428	0.840	0.840	0.840
4	84.5404%	0.6906	0.2759	0.3419	0.846	0.845	0.845
5	84.2618%	0.6851	0.2748	0.343	0.843	0.843	0.843
6	84.5404%	0.6907	0.2733	0.3404	0.845	0.845	0.845
7	84.4011%	0.6879	0.2768	0.3438	0.844	0.844	0.844
8	84.9582%	0.699	0.2741	0.3401	0.850	0.850	0.850
9	85.7939%	0.7157	0.2733	0.3398	0.858	0.858	0.858
10	84.6797%	0.6934	0.275	0.3405	0.847	0.847	0.847
Media	84.6797%	0.69343	0.27522	0.34166	0.8469	0.8468	0.8468
Desviación estándar	0.49%	0.009728	0.001380	0.001496	0.004841	0.004872	0.004871

Con el fin de ilustrar los resultados obtenidos en los demás algoritmos clasificadores, se generaron las figuras 3.8 y 3.9, las cuales muestran el comportamiento en cada parámetro considerado.

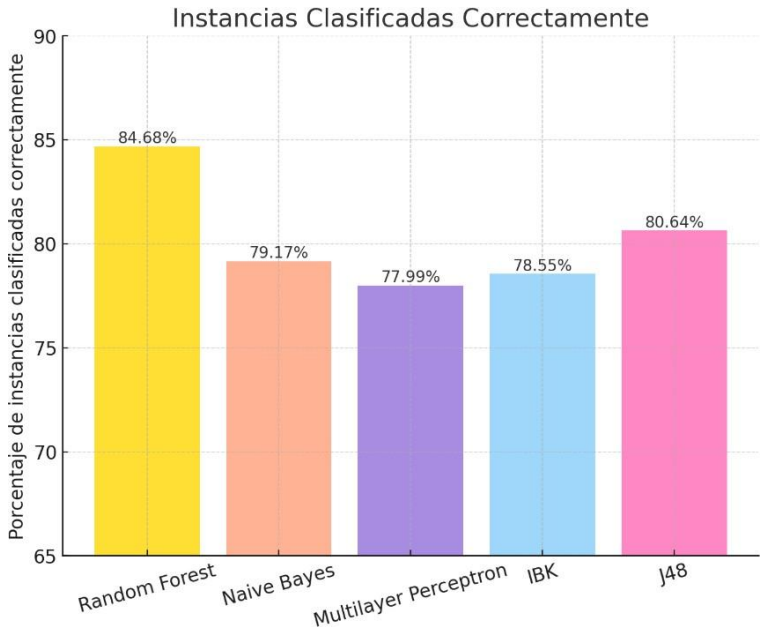

Figura 3.5 Gráfica de comparación entre algoritmos clasificadores para instancias clasificadas correctamente en el tercer modelo

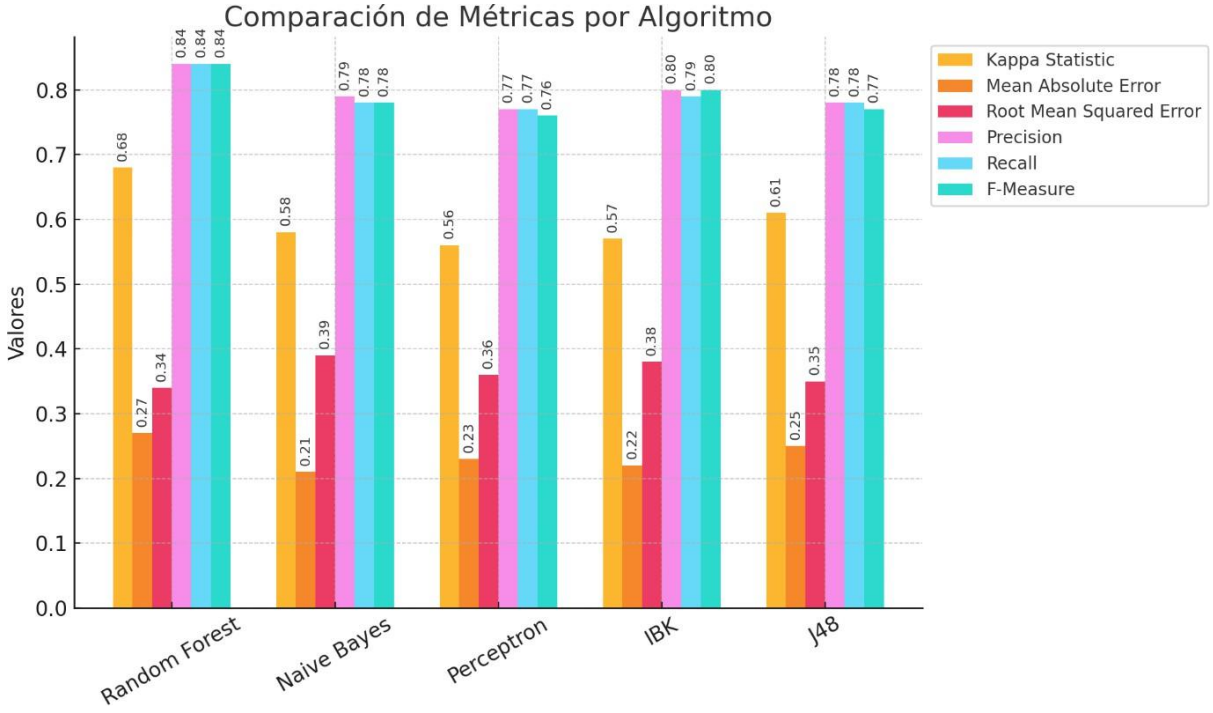

Figura 3.6 Gráfica de comparación entre algoritmos clasificadores para parámetros de error y precisión en el tercer modelo

3.2 Aplicación de metodología desarrollo

En esta sección, se explica la implementación de la metodología de desarrollo aplicada para el proyecto, la cual está basada en Scrum. Seleccionada por su enfoque iterativo y su capacidad para adaptarse a los cambios durante el proceso de construcción de la plataforma, la metodología Scrum responde a la necesidad de un desarrollo ágil que permita incorporar de manera flexible nuevas funcionalidades, impulsando así un crecimiento sostenido de la plataforma.

Dado que la plataforma médica pretende escalar a futuro cubriendo más enfermedades ginecológicas, es necesario implementar el diseño de una arquitectura escalable.

3.2.1 Pila de producto

La pila de producto representa una lista ágil de los requerimientos y funcionalidades que dirigen el desarrollo de la plataforma. Por tratarse de dos aplicaciones para el sistema, se definieron dos pilas por separado. En la tabla 3.15 se muestra la pila de producto de la aplicación web, tomando en cuenta la funcionalidad básica y el flujo de interacción inicial para los usuarios. En la primera sección se encuentran aquellos que son básicos para establecer el perfil del usuario y gestionar la interacción esencial con pacientes. Por otro lado, los casos de uso relacionados con diagnósticos e identificación de necesidades médicas se ubican en una prioridad de nivel medio, mientras que los casos de uso de administración, como la modificación de citas y horarios se consideran en una prioridad menor.

Tabla 3.15 Pila del producto de la aplicación web

ID	Descripción	Prioridad
1	Autorregistrarse	Alta
2	Verificar correo	Alta
3	Iniciar sesión	Alta
4	Cerrar sesión	Alta
5	Crear nuevo paciente	Alta
6	Aceptar solicitud de paciente	Alta
7	Agregar paciente a la agenda	Alta

ID	Descripción	Prioridad
8	Crear cita	Alta
9	Consultar agenda	Media
10	Consultar pacientes	Alta
11	Ver detalles de paciente	Alta
12	Realizar diagnóstico	Alta
13	Identificar necesidad de estudios de laboratorio	Alta
14	Identificar necesidad de especialista	Alta
15	Modificar datos personales	Media
16	Ver estadísticas	Media
17	Actualizar paciente	Alta
18	Modificar horario	Media
19	Modificar cita	Media
20	Eliminar cita	Media

De igual forma, se muestra la pila de producto de la aplicación móvil en la tabla 3.16.

Tabla 3.16 Pila del producto del sistema móvil

ID	Descripción	Prioridad
1	Autorregistrarse como paciente	Alta
2	Iniciar sesión	Alta
3	Cerrar sesión	Alta
4	Visualizar datos generales	Alta
5	Registrar datos generales	Alta
6	Actualizar datos generales	Alta
7	Visualizar datos ginecológicos	Alta
8	Registrar datos ginecológicos	Alta
9	Actualizar datos ginecológicos	Alta

ID	Descripción	Prioridad
10	Visualizar antecedentes	Alta
11	Registrar antecedentes	Alta
12	Actualizar antecedentes	Alta
13	Registrar periodo menstrual	Media
14	Calcular necesidad de visita médica	Alta
15	Visualizar cita médica	Alta
16	Contactar médico	Alta
17	Aceptar solicitud médico	Alta

De forma gráfica, en las figuras 3.7 y 3.8, se muestran los casos de uso tanto para las aplicaciones que forman la plataforma, estos diagramas permiten visualizar de forma clara las interacciones entre los usuarios y la aplicación.

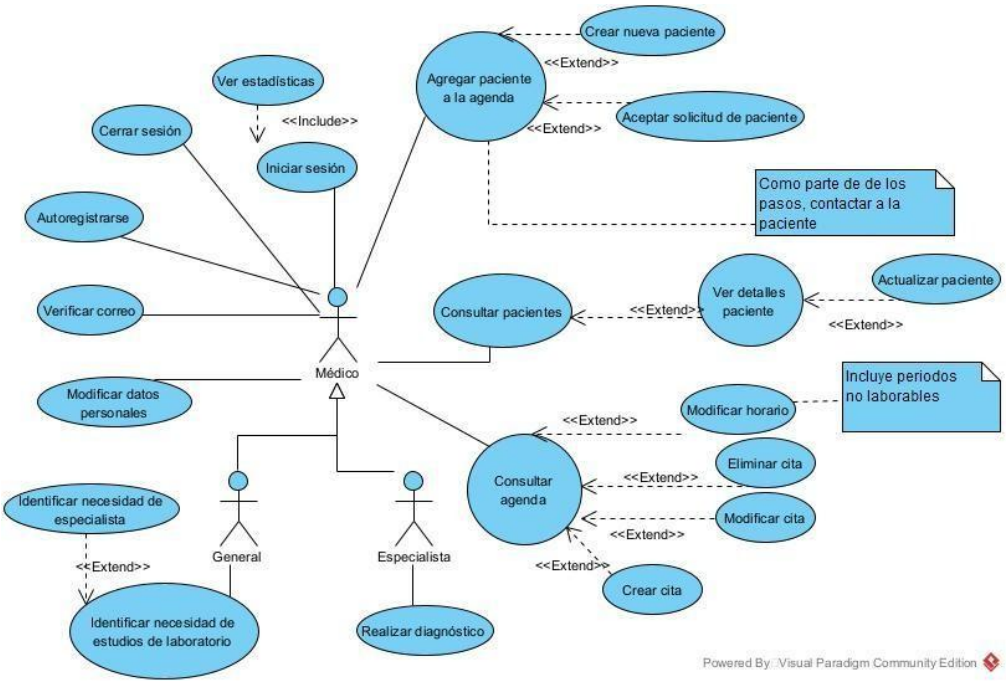

Figura 3.7 Diagrama de casos de uso de sistema Web

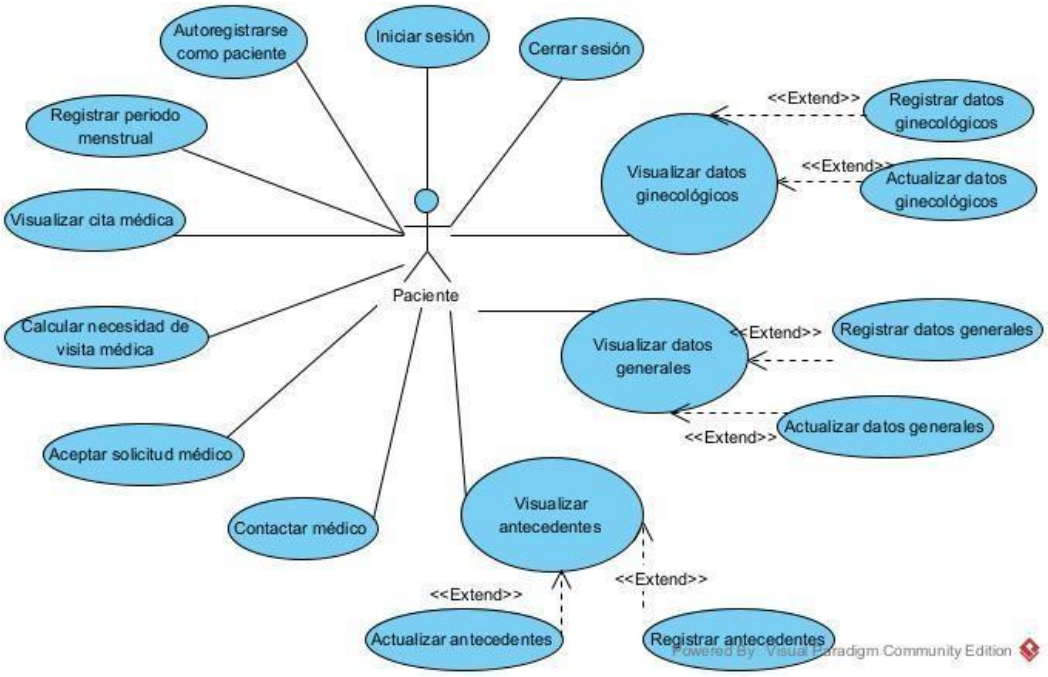

Figura 3.8 Diagrama de casos de uso aplicación móvil

Con el fin de ilustrar el funcionamiento de la plataforma y las necesidades requeridas por el usuario final, se describen a continuación algunos casos de uso relevantes. El resto de las descripciones de los casos de uso se encuentran en el Anexo II (archivo por separado, AnexoIICU.pdf).

Caso de Uso: Autorregistrarse (aplicación Web)

Objetivo: permitir a un médico registrarse en el sistema de forma autónoma.

Actor principal: Médico

Precondiciones:

- El médico no debe tener una cuenta existente en el sistema.

Postcondiciones:

- Se crea una cuenta nueva para el médico.
- Se envía un correo de verificación al médico.

Tabla 3.17 Flujo básico del Caso de Uso Autorregistrarse

Actor	Sistema
1. Selecciona la opción de registrarse.	
	2. Muestra un formulario de registro con los campos: nombre, primer apellido, segundo apellido, foto, teléfono, cédula profesional, dirección de su consultorio, estado, municipio, localidad, código postal, horario de atención, contraseña y confirmación de la contraseña.
3. Completa el formulario y envía los datos.	
	4. Verifica que todos los campos sean correctos. Si no lo son, ver Curso de Excepción "Error en el Formulario". 5. Todos los campos son requeridos, excepto el segundo apellido y la fotografía. El número de teléfono es de diez dígitos. La cédula profesional es de siete a ocho dígitos. El código postal es de cinco dígitos. La contraseña es de doce caracteres. El horario debe contener al menos inicio y fin de turno de un día. Los horarios de fin de turno deben de ser mayores que los horarios de inicio de turno. Si hay más de un horario de días, los días deben ser excluyentes.

Actor	Sistema
	6. Crea la cuenta como inactiva con fecha de registro el día actual y envía un correo de verificación (genera una URL con el correo y un token de seguridad). 7. Indica al médico que debe verificar su cuenta a través del correo enviado. 8. Termina el caso de uso.

Tabla 3.18 Curso de excepción "Error en el formulario" de Caso de Uso Autorregistrarse

Actor	Sistema
	1. Indica qué campos son erróneos o están incompletos. 2. Permite al médico corregir el formulario. 3. Termina el caso de uso.

Requerimientos no funcionales específicos: el sistema debe enviar correos de verificación y gestionar el almacenamiento seguro de contraseñas.

Pendientes: No hay

Supuestos considerados: No hay.

Caso de Uso: Iniciar Sesión (aplicación móvil)

Objetivo: permitir a la paciente autenticarse en la aplicación móvil de forma segura.

Actor principal: Paciente

Precondiciones:

- La paciente debe tener una cuenta registrada en el sistema.

Postcondiciones:

- La sesión de la paciente queda abierta en el sistema.
- Se registra un inicio de sesión exitoso.

Tabla 3.19 Flujo básico del Caso de Uso Iniciar Sesión

Actor	Sistema
1. Ingresa sus credenciales (Correo y contraseña)	
	2. Si las credenciales no son válidas, ver Curso de Excepción "Error de Autenticación". 3. Si las credenciales son válidas, abre la sesión de la paciente y redirige al panel principal. 4. Termina el caso de uso.

Tabla 3.20 Curso de excepción "Error de autenticación" de Caso de Uso Iniciar Sesión

Actor	Sistema
	1. Indica que las credenciales son incorrectas y permite intentarlo nuevamente. 2. Termina el caso de uso.

Requerimientos no funcionales específicos: no hay.

Pendientes: no hay.

Supuestos considerados: no hay.

También se muestran ejemplos del diseño preliminar de la interfaz de usuario de la plataforma, tanto en su versión web como en la móvil, permitiendo visualizar la disposición de algunos de los elementos principales en cada pantalla.

Dado que son múltiples pantallas generadas para el maquetado, se incluyen dos imágenes representativas de la aplicación web y dos de la aplicación móvil. En el caso de la aplicación web se muestran las pantallas de cálculo de SOP (Figuras 3.9 y 3.10) y de antecedentes clínicos.

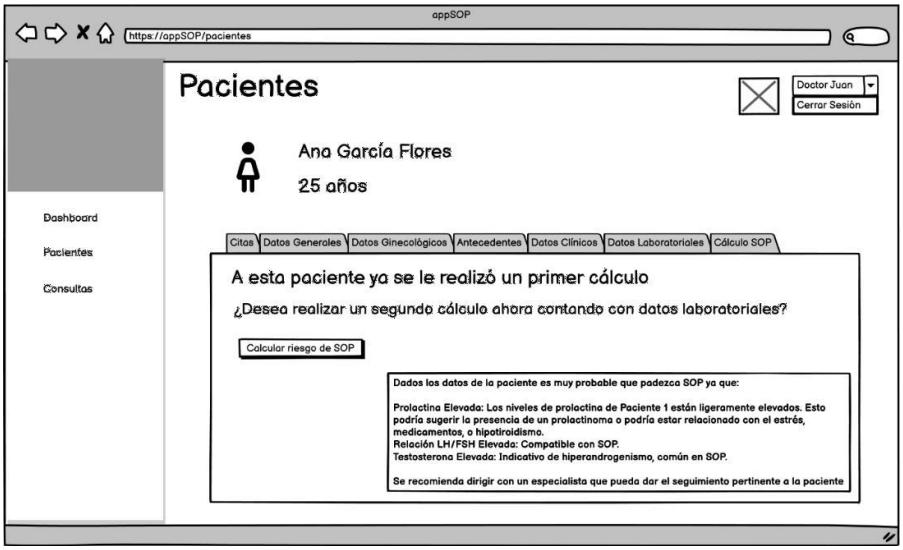

Figura 3.9 Maquetado sistema Web de la pantalla Cálculo SOP

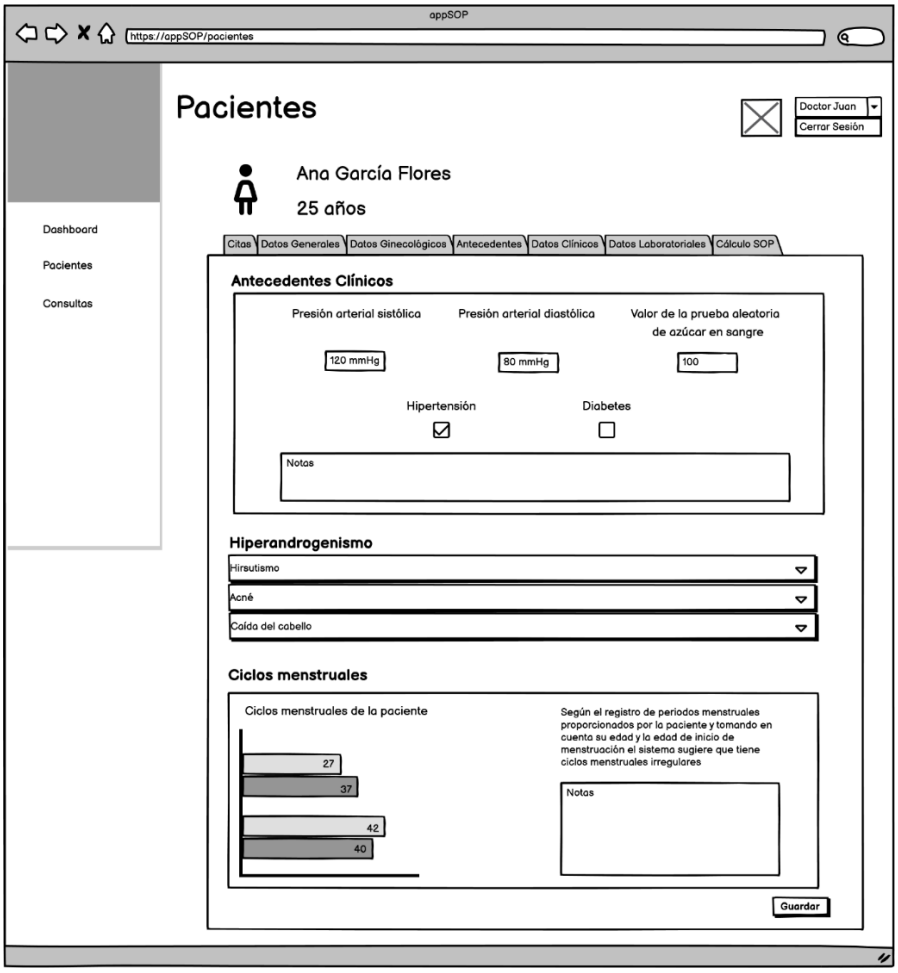

Figura 3.10 Maquetado sistema Web de la pantalla Datos Clínicos

Por otro lado, se tienen las pantallas de la aplicación móvil las cuales hacen referencia al menú y al cálculo de necesidad visita médica respectivamente (Figura 3.11). Estas imágenes se enfocan en los elementos de navegación y en la organización de las principales funcionalidades, lo cual permite una primera aproximación a la experiencia de usuario en ambas aplicaciones que conforman la plataforma.

Figura 3.11 Maquetado aplicación móvil de las pantallas Menú y Calcular Necesidad de Visita Médica

3.2.2 Artefactos de ingeniería de software

En este apartado se presentan los principales artefactos de ingeniería de software utilizados para la plataforma. Dichos artefactos se elaboraron utilizando la herramienta Visual Paradigm y los diagramas completos se incluyen como Anexo III en el archivo adjunto a esta tesis (AnexoIII_ArtefactosSw.vpp).

3.2.2.1 Diagrama de clases a nivel de análisis

Este diagrama presenta las clases que representan el dominio del negocio y las relaciones que existen entre ellas, enfocándose en la estructura conceptual de los sistemas web y móvil. Tal como lo muestra el diagrama en la figura 3.12, donde se ilustra una parte del diagrama de clases de la plataforma, donde se visualizan las clases referentes al paciente con sus atributos, el diagrama completo se encuentra en el Anexo III mencionado antes.

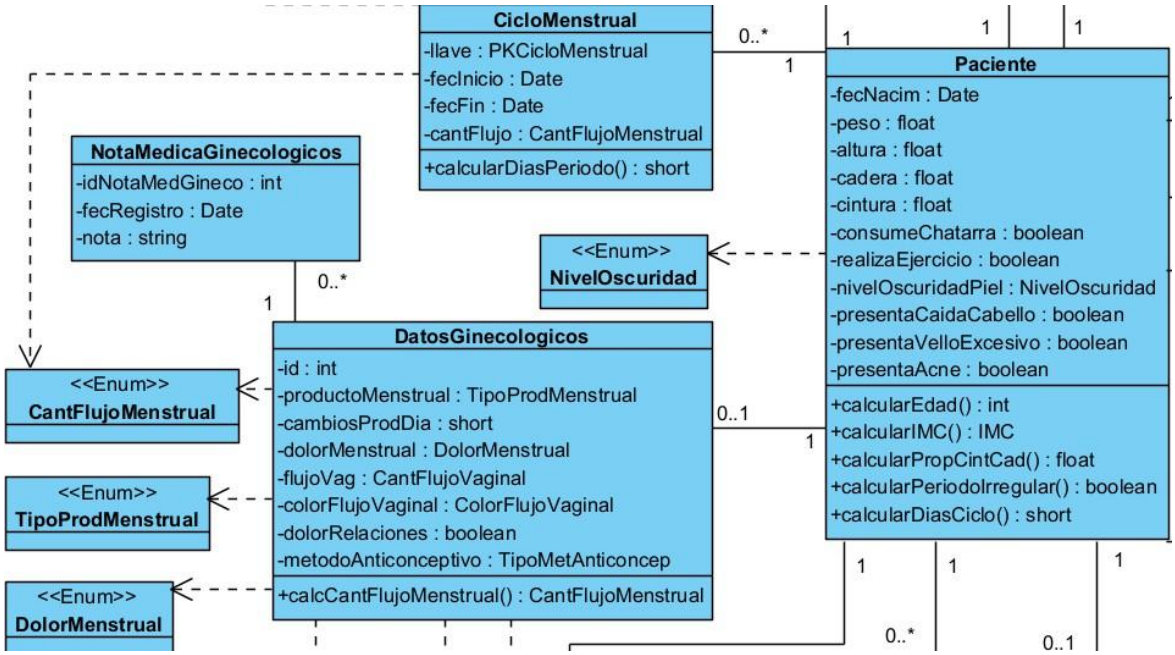

Figura 3.12 Diagrama parcial de clases a nivel de análisis

3.2.2.2 Diagrama de clases a nivel de diseño

A partir del análisis y de las características particulares de las tecnologías seleccionadas para el desarrollo (*back-end* Java, Weka, Spring; *front-end* JavaScript y React), fue necesario realizar ajustes al diagrama de clases original.

Java utiliza los valores enumerados para facilitar la lectura de código y éste quede autodocumentado en el caso de características discretas como el nivel de acné o la caída del cabello, de modo que se agregaron clases con el estereotipo `Enum` para distinguir estos casos.

Por otra parte, el módulo de Spring relacionado con el mapeo objeto-relacional (programación orientada a objetos cuya representación de persistencia es una base de datos relacional) tiene requerimientos específicos para ciertas condiciones; por ejemplo, en la base de datos es posible que existan llaves primarias formadas de varios campos, mientras que en Spring sólo es posible anotar como llave un único atributo de la clase, por lo que se hace necesario incluir clases especiales que agrupan los campos de la llave en la base y la clase original incluye ahora un atributo de la clase especial. En la figura 3.13 se muestra un fragmento del diagrama de clases a nivel diseño, donde aparecen valores enumerados y la clase `PKDatosLaboratoriales` que agrupa a los atributos correo (*email*) y fecha de estudio que funcionan como llave en la base de datos para la tabla que almacena los Datos Laboratoriales de la paciente.

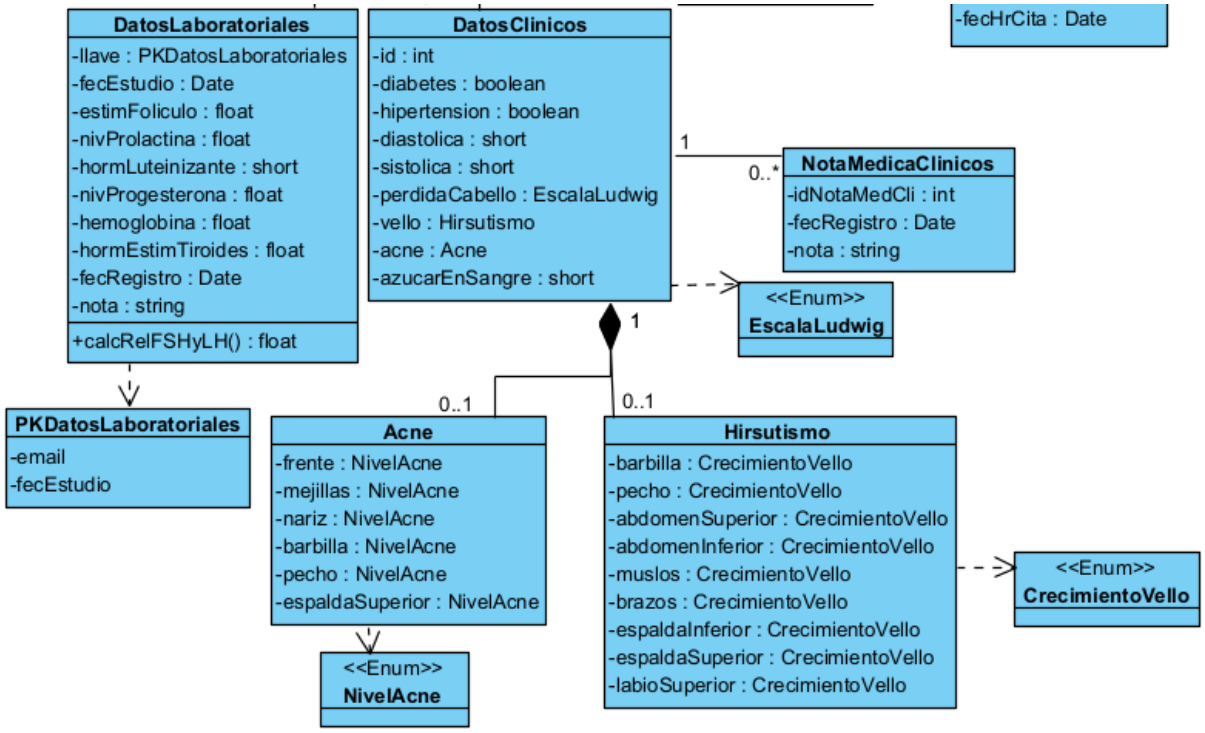

Figura 3.13 Diagrama de clases a nivel de diseño

3.2.2.3 Diagrama de Estructura Física de Base de Datos

Este diagrama representa la estructura de almacenamiento de los datos gestionados por la plataforma, detallando tablas, relaciones y llaves que aseguran la integridad y el correcto

funcionamiento de la información. Se muestra parte de la sección relacionada con el paciente (figura 3.14).

La estructura de la base se vio afectada por las decisiones tomadas para el diagrama de clases debido a la naturaleza de Spring como se mencionó en el apartado anterior. En algunos casos se decidió crear campos específicos para las llaves primarias en lugar de utilizar llaves compuestas, esto para evitar la proliferación de clases extra.

El diagrama realizado permitió la obtención del código fuente del script de creación de la base de datos relacional.

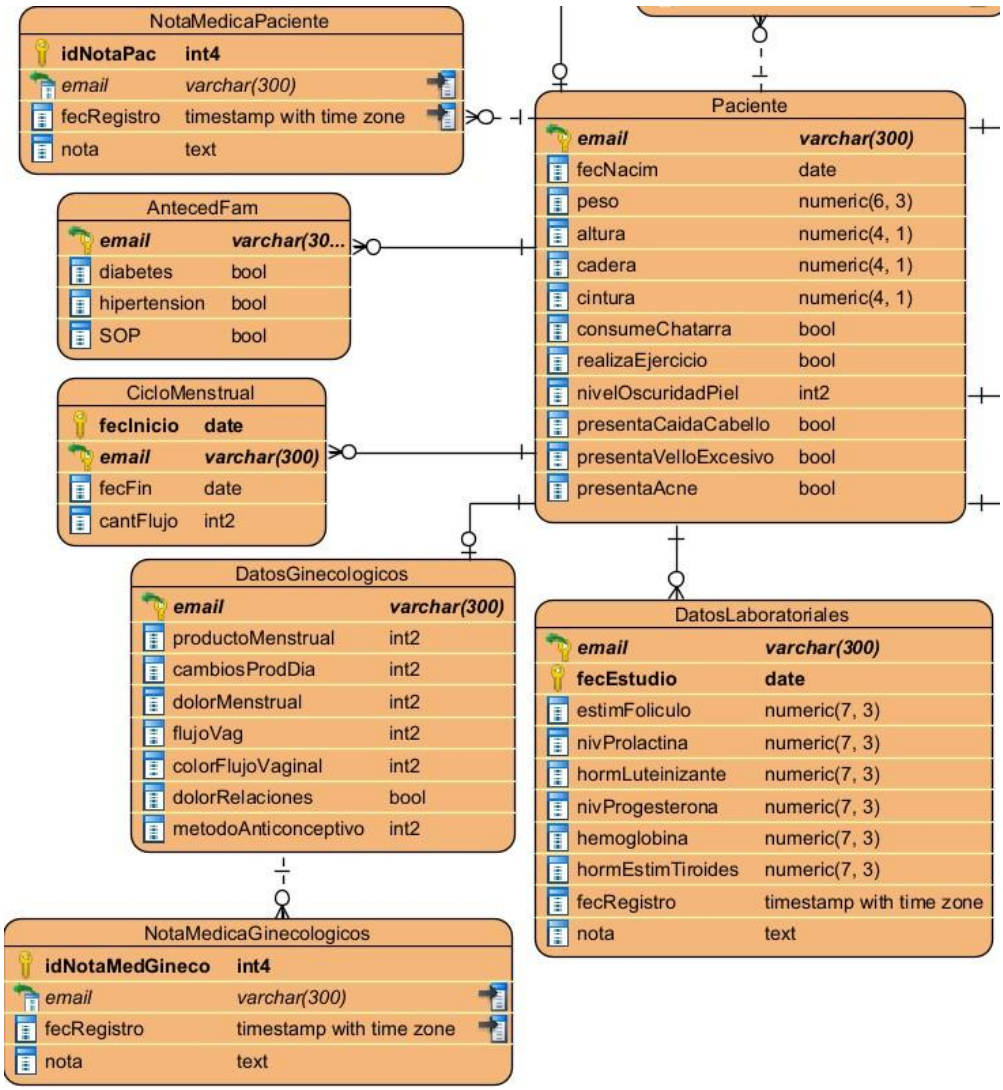

Figura 3.14 Diagrama de Base de Datos

3.2.2.4 Diagrama de Paquetes

El diagrama de paquetes (figura 3.15) organiza los distintos paquetes que componen el *back-end* de la plataforma y se ve afectado por las características propias de la tecnología utilizada, en este caso Spring.

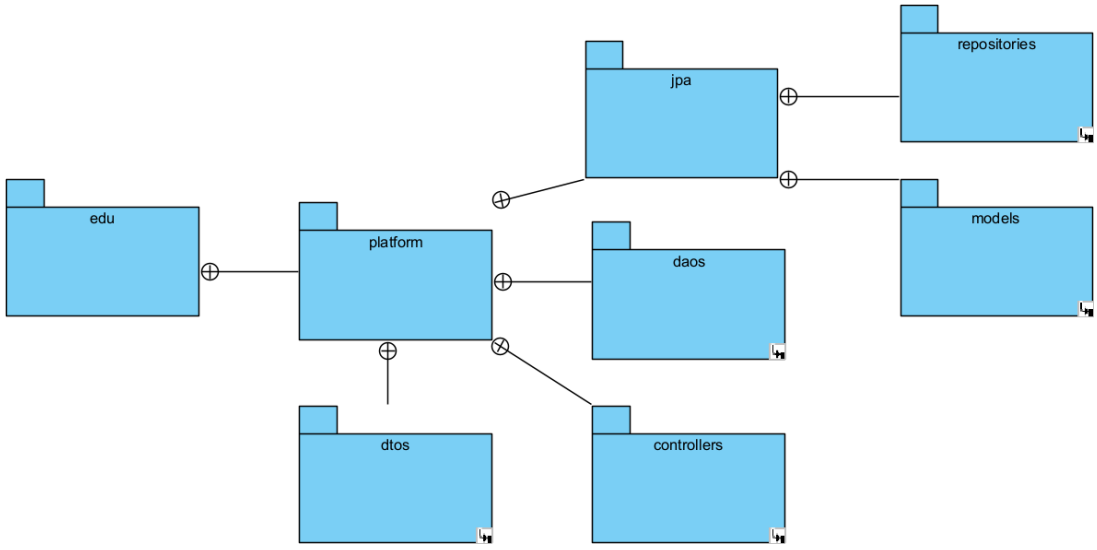

Figura 3.15 Diagrama de paquetes

- Los paquetes `edu` y `platform` están relacionados únicamente con la estructura que formará la dirección URL del API.
- El paquete `jpa` corresponde a la parte que controla la persistencia de la información. Debido a las características de Spring, se divide en dos paquetes.
    - En el paquete `models` se encuentran las clases que representan el dominio del negocio y tienen persistencia directa en la base de datos.
    - Por otro lado, en el paquete `repositories` se definen las interfaces que extienden de `CrudRepository`, una interfaz que provee Spring y que permite realizar consultas, inserciones, actualizaciones y eliminaciones en la base de datos relacional sin necesidad de escribir código SQL.
- El paquete `daos` contiene las clases que implementan las reglas de negocio y las transacciones especiales. Estas clases son las únicas que se comunican con los elementos del paquete `jpa`.

- En el paquete `controllers` se encuentran las clases que implementan el API del *back-end*. Éstas hacen uso de las clases contenidas en el paquete `daos`.
- Para ocultar las condiciones particulares del mapeo objeto-relacional, el paquete `dtos` contiene clases simples, formadas únicamente por atributos. Estas clases permiten la conversión entre objetos JSON y clases Java, las utilizan tanto los elementos del paquete `controllers` como los del paquete `daos`.

Es importante hacer notar que, gracias al desarrollo de los diagramas de clases y paquetes, se obtuvo el código básico para la construcción de la plataforma de forma automática.

3.2.2.5 Prototipo de navegación web

A partir del maquetado original, se realizó un prototipo de navegación en React para la versión web, el cual ilustra el flujo de pantallas y la disposición de elementos interactivos dentro de ellas, facilitando la visualización de la experiencia de usuario en la versión de escritorio del sistema.

Ya que son múltiples pantallas generadas para el prototipo de navegación, se incluyen sólo dos imágenes representativas del diseño de la aplicación web donde se muestra pantalla de citas de una paciente (Figura 3.16) y pantalla de agenda del médico (Figura 3.17).

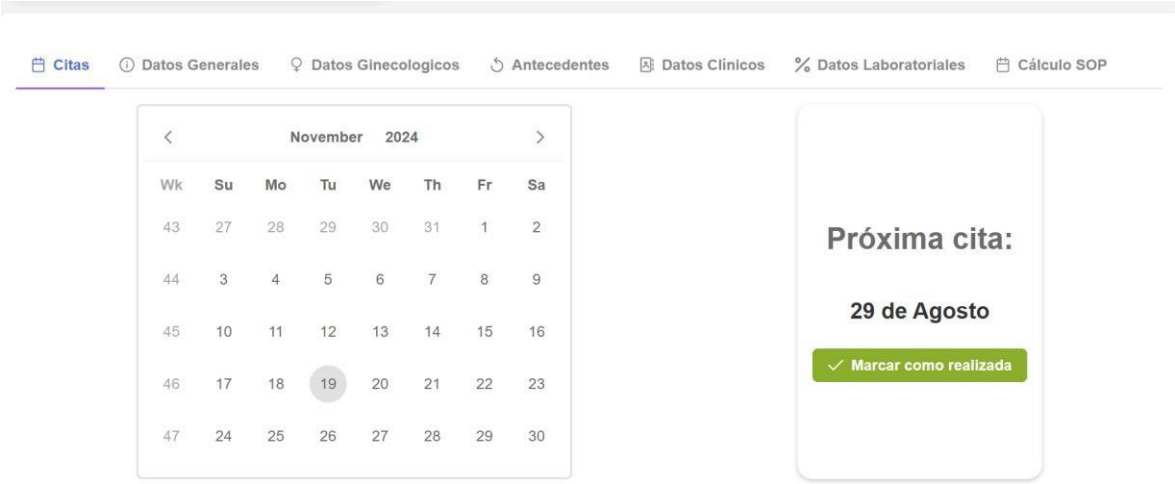

Figura 3.16 Prototipo de pantalla de Citas de una paciente

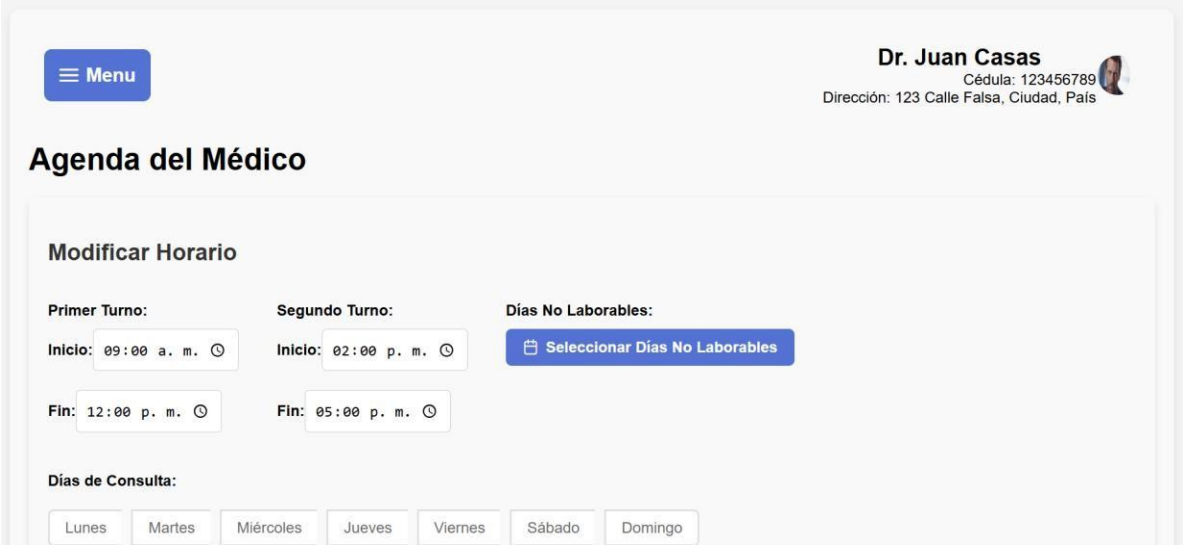

Figura 3.17 Prototipo de pantalla de Agenda del Médico

3.2.2.6 Prototipo de navegación móvil

También se realizó el prototipo de navegación para la versión móvil de la plataforma mediante React Native, mediante éste es posible visualizar el flujo de trabajo y la disposición de algunos elementos interactivos con el fin de ilustrar la experiencia del usuario, en este caso la paciente, para validarla con la médico general. Se muestra las pantallas de menú y de datos ginecológicos de la app (Figura 3.18).

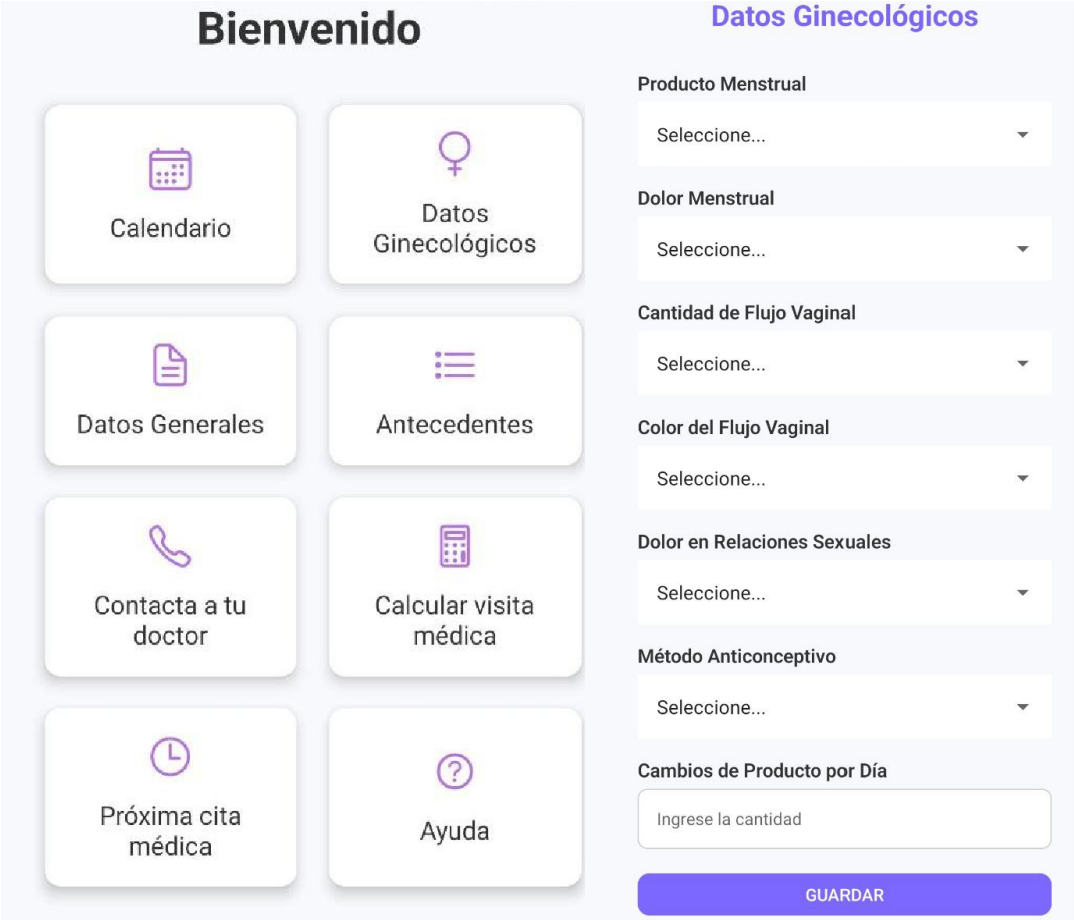

Figura 3.18 Prototipo de pantallas de Menú y Datos Ginecológicos en aplicación móvil

3.2.3 Desarrollo de aplicaciones

Con el objetivo de asegurar una implementación estructurada de la plataforma propuesta, se implementaron entregas incrementales denominadas *sprints*, debido a que se adoptó la metodología ágil SCRUM para este proyecto. Este enfoque favoreció una evolución continua del sistema.

En las siguientes secciones se describe de manera abreviada cada uno de los `sprints` realizados; debido a que ambas aplicaciones (app y web) utilizan el mismo *back-end*, se consideró apropiado mostrar código sólo de éste para facilitar la lectura del documento.

3.2.3.1 Sprint 1 Prueba de envío de correos desde el *back-end*

Objetivo: Implementar el servicio de envío de correos electrónicos desde el *back-end* para su uso en el proceso de verificación de cuenta.

Actividades realizadas:

- Configuración de `JavaMailSender` en Spring Boot.
- Creación de la clase `EmailService`, con métodos para enviar correos de texto plano y HTML.
- Pruebas con cuentas SMTP configuradas para entorno local.
- Generación de enlaces de verificación mediante tokens únicos.

Un fragmento del código implementado para esta sección se muestra en el listado 1; en las líneas 1 y 2 se crean los objetos necesarios para el envío de correo, en las líneas 3 a 5 se llenan los datos del correo (a quién va dirigido, el título y el contenido del mensaje) mientras que en la línea 6 se envía el mensaje una vez configurado.

Listado 1 Código de verificación de correo

Lín.	Código
1	`MimeMessage message = mailSender.createMimeMessage();`
2	`MimeMessageHelper helper = new MimeMessageHelper(message, true);`
3	`helper.setTo(email);`
4	`helper.setSubject(subject);`
5	`helper.setText(content, true);`
6	`mailSender.send(message);`

3.2.3.2 Sprint 2 Registro de paciente y médico

Objetivo: Desarrollar el API para el proceso de registro de usuarios (paciente y médico) en la plataforma, incluyendo validaciones, almacenamiento y notificaciones.

Actividades realizadas para el *back-end*:

- Implementación de los métodos de las clases de modelo `Paciente` y `Médico` en el paquete `jpa`.
- Construcción de las clases:
    - `Paciente, PacienteDTO, PacienteDAO`
    - `Medico, MedicoDTO, MedicoDAO`
    - `Usuario, UsuarioDTO, UsuarioDAO`
    - `AuthController`
    - `LoginRequest`
    - `LoginResponse`
    - `ApiClient`
- Implementación de los métodos `register` (para el caso de pacientes) y `registerMed` (para el personal médico) de la clase `AuthController`, con validación de correo duplicado y envío de correo de confirmación
- Uso de API externa proporcionada por la plataforma Rapid Api, para la verificación de cédula profesional del médico.

El listado 2 contiene un método de la clase `ApiClient` que hace uso del API de cédulas profesionales de RapidApi; dentro de la línea 1 se establece que se recibirá un parámetro de tipo cadena que contiene la cédula profesional a revisar y dentro de la línea 3 se coloca dicha cédula dentro de la URI a enviar, se espera la respuesta que se evalúa dentro de la condicional de la línea 8 y se prepara la propia respuesta del método; en caso de ser exitosa la búsqueda (línea 9) o de lo contrario el error a devolver (línea 11).

Listado 2 Código de implementación de API externa

Lín.	Código
1	`public ApiResponse fetchApiData(String cedProf) throws IOException, InterruptedException {`
2	`HttpRequest request = HttpRequest.newBuilder()`
3	`.uri(URI.create(API_URL + "?cedula=" + cedProf))`
4	`.header("x-rapidapi-key", API_KEY)`

Lín.	Código
5	`.header("X-RapidAPI-Host","cedulas-profesionales sep.p.rapidapi.com")`
6	`.GET()`
	`.build();`
7	`HttpResponse<String> response = httpClient.send(request, HttpResponse.BodyHandlers.ofString());`
8	`if (response.statusCode() == 200) {`
9	`return objectMapper.readValue(response.body(), ApiResponse.class);`
10	`} else {`
11	`throw new IllegalArgumentException("Error al llamar al API. Código de respuesta: " + response.statusCode());`
12	`}`
13	`}`
14	`}`

Actividades realizadas para *front-end*:

- Desarrollo de formulario de registro en React Native para el caso de la app (paciente) y en React para el caso web (médico).

3.2.3.3 Sprint 3 Inicio de sesión

Objetivo: Implementar el proceso de autenticación mediante JWT, diferenciando usuarios por roles y asegurando la persistencia de sesión dentro de la aplicación.

Actividades realizadas para el *back-end*:

- Validación de credenciales con `Bearer`.
- Generación de token JWT con fecha de expiración.
- Envío de token en respuesta al *login* exitoso.

En el listado 3 se presenta un fragmento del código de la clase `AuthController`, que utiliza las ventajas que proporciona Spring en su módulo de seguridad. La anotación de la línea 1 indica que la llamada al API se hará mediante el método POST; en la línea 2 se indica que en el

contenido del cuerpo del mensaje recibido se encuentra un objeto de tipo `LoginRequest` (un DTO que contiene el correo electrónico y la contraseña para ingresar) y que el método tiene varias posibles respuestas. En la línea 4 se invoca al método `validateUser` de un objeto de la clase `UsuarioDAO` (este objeto se obtiene mediante la inyección de dependencias propia de Spring), el nombre y firma de tal método lo define Spring en su módulo de seguridad y en la línea 5 se genera el *token* de seguridad, también bajo las condiciones propias de Spring.

En caso de éxito, se devuelve un objeto de tipo `LoginResponse` (también es un DTO) que incluye el token generado como se ve en la línea 6. Las otras posibles respuestas se producen porque no se reconocen las credenciales recibidas (líneas 7 a 10) o porque se produce algún tipo de error durante la ejecución (líneas 11 a 13).

Listado 3 Código de ingreso (login)

Lín.	Código
1	`@PostMapping("/login")`
2	`public ResponseEntity<?> login(@RequestBody LoginRequest loginRequest) {`
3	`try {`
4	`Usuario usuario = usuarioDAO.validateUser(loginRequest.getEmail(), loginRequest.getPassword());`
5	`String token = jwtService.generateToken(usuario);`
6	`return ResponseBuilder.success("Login exitoso", new LoginResponse(token));`
7	`} catch (RuntimeException e) {`
8	`log.warn("Login fallido para email: {}", loginRequest.getEmail());`
9	`return ResponseBuilder.unauthorized( "Credenciales inválidas.");`
10	`} catch (Exception e) {`
11	`log.error("Error inesperado en login: ", e);`
12	`return ResponseBuilder.error("Error inesperado al intentar iniciar sesión.");`
13	`}`

Lín.	Código
14	}

Actividades realizadas para el *back-end*:

- Desarrollo de formulario de inicio de sesión para app y para aplicación Web.
- Almacenamiento del token en `AsyncStorage` para mantener la sesión.
- Redirección condicional al menú/tablero del usuario autenticado.

3.2.3.4 Sprint 4 Registro de datos complementarios de la paciente

Objetivo: Permitir la captura de información clínica completa del paciente, incluyendo datos generales, ginecológicos, antecedentes personales y heredofamiliares.

Debido a la gran cantidad de datos que se capturan de la paciente, cada sección se almacena de forma independiente, para facilitar la interacción y evitar recapturas en caso de errores, por lo que existen controladores independientes para cada sección.

Actividades realizadas para el *back-end*:

- Implementación de métodos en clases de modelo `DatosGenerales`, `DatosGinecologicos`, `AntecedGineco` y `AntecedFam`.
- Creación de controladores para guardar y consultar datos por sección de paciente.

En el listado 4 se muestra un fragmento de código de la clase AntecedFamController, específicamente el método que permite la eliminación de los datos. En la línea 1 se encuentra la anotación que indica que esta parte del API responde al método DELETE de HTTP y que, como parte de la URL, recibe el identificador de los antecedentes familiares (líneas 1 y 2). Las líneas 3 a 5 muestran la validación previa que se realiza de los datos y la respuesta en caso de no pasar dicha validación, mientras que las líneas finales contienen el código para la eliminación y la generación de respuesta correspondiente.

Listado 4 Código de eliminación de antecedentes familiares

Lín.	Código
1	`@DeleteMapping("/{id}")`
2	`public ResponseEntity<?> eliminar(@PathVariable int id) {`
3	`    if (antecedFamDAO.obtenerPorId(id).isEmpty()) {`
4	`      return ResponseBuilder.notFound("Registro no encontrado para eliminar.");`
5	`    }`
6	`    antecedFamDAO.eliminar(id);`
7	`    return ResponseBuilder.success("Registro eliminado correctamente.");`
8	`}`

Actividades realizadas para el *front-end*:

- Desarrollo de componentes para cada tipo de formulario.
- Validación de datos, control de cambios no guardados y alertas visuales.

3.2.3.5 Sprint 5 Gestión de la agenda médica

Objetivo: Implementar la funcionalidad de gestión de citas médicas, permitiendo a la paciente mantenerse al tanto de sus citas y al médico organizar su agenda dentro del sistema.

Actividades realizadas para el *back-end*:

- Implementación de clase de modelo Cita.
- Implementación de *controllers* para consultar citas por paciente y filtrar por el estatus en el que se encuentre (realizada o no realizada).
- Realización de modificaciones en la cita.

En el listado 5 se muestra un fragmento del código de la clase `CitaController`, específicamente el método que consulta las citas de la agenda de un médico. En la línea 1 la anotación indica que se trata de una consulta (`GetMapping`) y que como parte de la URL se espera el identificador del médico, mismo que se tomará como parámetro de búsqueda (líneas 1

y 2). En las líneas 3 a 7 se verifica que exista el médico con el correo recibido como parámetro; en la línea 8 se realiza la consulta de las citas y se genera una respuesta específica si no hay citas para el médico indicado (líneas 9 a 11) o se devuelve el conjunto de citas encontradas (línea 12).

Listado 5 Código de eliminación de antecedentes familiares

Lín.	Código
1	`@GetMapping("/medico/{email}")`
2	`public ResponseEntity<?> obtenerCitasPorMedico(@PathVariable String email) {`
3	`    try {`
4	`        medicoDAO.obtenerMedicoPorEmail(email);`
5	`    } catch (RuntimeException e) {`
6	`        return ResponseBuilder.notFound("Médico no encontrado");`
7	`    }`
8	`    List<Cita> citas = citaDAO.obtenerCitasPorMedico(email);`
9	`    if (citas.isEmpty()) {`
10	`        return ResponseBuilder.notFound("No hay citas correspondientes al médico");`
11	`    }`
12	`    return ResponseBuilder.success("Citas encontradas", citas);`
13	`}`

Actividades realizadas para el *front-end*:

- Visualización de citas en formato calendario (app y web).
- Diálogo modal de detalles al seleccionar una cita (app y web).
- Leyenda para identificar el estado de cada cita (app y web).
- Crear cita utilizando la interfaz de calendario y reloj para especificar los detalles (web).
- Modificar cita reubicándola dentro del calendario (web).

3.2.3.6 Sprint 6 Ejecución de modelos predictivos

Objetivo: Implementar la funcionalidad para utilizar los tres modelos de clasificación realizados con KDD.

Actividades realizadas para el *back-end*:

- Implementación de la clase `PrediccionController` que tiene la capacidad de ejecutar los tres modelos de acuerdo a la información disponible.
- Implementación de mensajes explicativos en lenguaje natural para facilitar la comprensión del resultado por parte de la paciente y del médico.
- Uso de clases auxiliares y servicios específicos para cada tipo de modelo.

La clase abstracta `ClasificacionService` define los elementos comunes para cualquier modelo de clasificación que se implemente en la plataforma, básicamente de dónde toma los mensajes explicativos, el clasificador y la instancia a clasificar de acuerdo a Weka, así como el método abstracto que deberá implementarse para el uso de cada modelo específico. En el listado 6 se encuentra el código de la clase mencionada.

En la línea 1 la anotación indica que la clase (y sus subclases) brinda servicios a otras y que dichos servicios no tienen un uso particular, a diferencia de los controladores (que forman el API) o de los modelos (que representan el mapeo a la base de datos). La línea 2 configura la relación con el archivo de propiedades que contiene los mensajes informativos en lenguaje natural. Las líneas 4 y 5 definen los atributos que se relacionan con Weka; la línea 6 define la firma del método que realizará la clasificación utilizando los modelos de minería de datos creados con KDD; como se ve en dicha firma, se requiere el objeto Paciente para realizar la clasificación y se devuelve PrediccionDTO (incluye la clase/predicción obtenida y, en su caso, las razones del resultado).

Listado 6 Código de eliminación de antecedentes familiares

Lín.	Código
1	`@Service`
2	`@PropertySource("classpath:/weka_models/textos_para_modelos.properties")`
3	`public abstract class ClasificacionService {`
4	`    protected Instances datos;`
5	`    protected Classifier clasificador;`

Lín.	Código
6	`public abstract PrediccionDTO clasifica(Paciente pac) throws Exception;`
7	`}`

En el listado 7 se encuentra un fragmento del código del método `clasifica`, de la clase `IrAlMedicoService`, misma que hereda de la clase `ClasificacionService` y que utiliza el primer modelo obtenido en KDD, es decir, calcula la necesidad de sugerir a la paciente que acuda a una visita médica si sus datos resultan anormales. En la línea 1 se indica que se está implementando el método heredado; en las líneas 3 a 7 se definen variables locales a utilizar; en la línea 8 se invoca al método que calcula los días de duración tanto del periodo como del ciclo (se basa en el historial de menstruación de la paciente y, si no existe, en los datos que ella haya indicado como antecedentes). En la línea 9 se crea la instancia a clasificar, indicando como tamaño el índice del último atributo (se suma 2 porque también el valor de clasificación ocupa espacio y la numeración del índice inicia en cero). En las líneas 10 a la 22 se pasan los valores de la información de la paciente a la instancia de Weka, haciendo uso de constantes para facilitar la lectura del código.

En la línea 24 se aplica el modelo entrenado sobre la instancia recién creada, mientras que en la línea 25 se obtiene la distribución de probabilidades resultante con lo que se conoce la certeza del modelo para la clasificación particular (línea 27). El primer modelo tiene dos posibles valores de clase, 0 para el caso de que no requiera ir al médico y 1 para el caso de que sí lo requiera, por lo que en la línea 28 se verifica si no se requiere asistir entonces arma la respuesta con un texto tranquilizador (línea 29); por otra parte, si se identificó que requiere atención médica, se arma la explicación de acuerdo a sus condiciones específicas (líneas 31 a 47) utilizando los textos en lenguaje natural sugeridos por la ginecóloga.

Listado 7 Código para aplicar el primer modelo

Lín.	Código
1	`@Override`
2	`public PrediccionDTO clasifica(Paciente pac) throws Exception{`
3	`    PrediccionDTO resultado;`
4	`    Instance porClasificar;`

Lín.	Código
5	double valor;
6	double[] distProba;
7	String razones;
8	pac.calcularDiasPeriodoyCiclo();
9	porClasificar = new DenseInstance(ATTR_SOP_EN_FAMILIARES+2);
10	porClasificar.setValue(datos.attribute(ATTR_GRUPO_EDAD), pac.calcularRangoEdad());
11	porClasificar.setValue(datos.attribute(ATTR_OSCURIDAD_PIEL), pac.getNivelOscuridadPiel().getNivel());
12	porClasificar.setValue(datos.attribute(ATTR_DOLOR_MENSTRUACION), pac.getDatosGinecologicos(). getDolorMenstrual().equals(DolorMenstrual.DURANTE)?1:0);
	…
23	datos.add(porClasificar);
24	valor = clasificador.classifyInstance(datos.lastInstance());
25	distProba = clasificador.distributionForInstance(datos.lastInstance());
26	resultado = new PrediccionDTO();
27	resultado.setCerteza(distProba[(int)valor]);
28	if (datos.classAttribute().value((int)valor).equals("0")){
29	resultado.setExplicacion(TXT_OK);
30	}else{
31	razones = RAZON_INICIO;
32	if (pac.isPeriodoIrregular())
33	razones =razones + "\n" + RAZON_IRREGULAR;
	…
46	razones = razones + "\n" + RAZON_FIN;
47	resultado.setExplicacion(razones);
48	}
49	return resultado;
50	}

La explotación de los modelos para identificar si es necesario solicitar exámenes laboratoriales y si es necesario referir a la paciente con un especialista por tener alta probabilidad de SOP, se

implementó de forma semejante a la mostrada para el primer modelo en las clases `PedirLaboratorialesService` e `IdentificarSOPService` respectivamente.

Actividades realizadas para el *front-end*:

- Implementación de llamada al `endpoint` de cada controlador (1 para la app y 2 para web) y presentación de resultado.

3.2.3.7 Sprint 7 Estadísticas

Objetivo: implementar un tablero (*dashboard*) en la aplicación web, que permita visualizar indicadores clave sobre la actividad de pacientes y citas médicas. Estas estadísticas tienen el propósito de brindar una vista general del uso del sistema y apoyar en la planificación de la atención médica.

Actividades realizadas para el *back-end*:

- Implementación de procedimientos en PostgreSQL para:
  - Contar pacientes por fecha de registro.
  - Obtener citas filtradas por fecha y situación.
- Creación de controladores
  - dashboard/resumen-citas-hoy.
  - dashboard/citas-agendadas-por-dia.
  - dashboard/pacientes-por-dia.
  - dashboard/total-pacientes.

Actividades realizadas para el *front-end*:

- Utilización de biblioteca Chart.js para generación de gráficas que permitan la visualización de los datos que envía el *back-end.*

En el siguiente capítulo, se describen los resultados obtenidos tras la validación del sistema, así como el impacto funcional y visual de su integración en los diferentes módulos.

Capítulo 4 Resultados

Como resultado de este proyecto de tesis se desarrolló una plataforma de salud denominada "Tepakilis Cihuatl", la cual se encuentra conformada por una aplicación web para médicos tanto generales como especialistas y una aplicación móvil orientada a pacientes.

Esta plataforma se apoya en técnicas de inteligencia artificial, específicamente utilizando minería de datos para identificar posibles casos de Síndrome de Ovario Poliquístico (SOP) mediante la recopilación, análisis y evaluación de datos clínicos, ginecológicos y laboratoriales de la paciente. El sistema integra tres modelos de clasificación, cada uno diseñado para apoyar la toma de decisiones médicas en diferentes etapas: sugerencia de consulta con un médico general, necesidad de solicitar exámenes clínicos y posibilidad de necesitar tratamiento con un médico especialista.

En el presente capítulo se muestran los resultados obtenidos tras la implementación y validación de la plataforma, incluyendo pruebas con casos de estudio reales. De igual manera, se muestran ejemplos del funcionamiento de las aplicaciones web y para dispositivo móvil.

Para acceder a las funcionalidades, la paciente debe iniciar sesión desde la aplicación móvil, mientras que el/la médico lo hace desde la aplicación web. Una vez dentro del sistema, el/la médico tiene la posibilidad de visualizar el estado general de salud ginecológica de la paciente y complementar los datos clínicos, de tal manera que el sistema analice dicha información y muestre al médico recomendaciones sobre la necesidad o no de solicitar exámenes laboratoriales a la paciente o evaluar la necesidad de una derivación a ginecología. En las siguientes secciones se describen los casos de estudio y sus resultados, así como el comportamiento tanto de la aplicación para dispositivo móvil como de la aplicación web.

4.1 Casos de estudio

El sistema implementa tres modelos de minería de datos, el primer modelo de clasificación se evalúa dentro de la aplicación móvil, tomando como referencia los datos ingresados por la usuaria, su objetivo es sugerir si la información reportada muestra anomalías que justifiquen una visita médica. Por su parte, en la aplicación web, se evalúa la ejecución del segundo y tercer modelo de clasificación, los cuales se encargan de apoyar al médico general en la toma de

decisiones respecto a la solicitud de exámenes clínicos y la derivación a un(a) especialista ante el riesgo de que la paciente padezca SOP respectivamente. Esto lleva a la necesidad de contar con datos de dos pacientes, una paciente sana para la que sólo se ejecuta el primer modelo y una paciente enferma, que ocupará los tres modelos.

En las siguientes tablas (4.1 a 4.4) se ilustran los datos a considerar (ginecológicos, generales, antecedentes gineco-obstétricos y antecedentes familiares) para los casos de prueba, haciendo énfasis en el hecho de que en los casos de estudio 2, 3 y 4 se trata de la misma paciente.

Tabla 4.1 Datos ginecológicos de casos de estudio

Datos ginecológicos	Caso de estudio 1	Casos de estudio 2, 3 y 4
Producto menstrual	Toalla sanitaria	Toalla sanitaria
Cambios de producto por día	3	7
Dolor menstrual	Sin dolor	Antes y durante
Cantidad de flujo vaginal	Escaso	Abundante
Color de flujo vaginal	Incoloro	Grisáceo
Dolor en relaciones sexuales	No	Sí
Método anticonceptivo	No utiliza	Condón

Tabla 4.2 Datos generales de casos de estudio

Datos generales	Caso de estudio 1	Casos de estudio 2, 3 y 4
Fecha de nacimiento	16/04/1998	03/09/1984
Peso	60 kg	68 kg
Altura	1.7 m	1.60 m
Índice de masa corporal (IMC)	20.76	26.56
Tamaño de la cadera	98 cm	105 cm
Tamaño de la cintura	72 cm	88 cm
Crecimiento de vello	No presenta	Sí presenta
Oscurecimiento en la piel	No presenta	Sí presenta

Datos generales	Caso de estudio 1	Casos de estudio 2, 3 y 4
Caída del cabello	No presenta	Sí presenta
Problemas de acné	No presenta	Sí presenta
Consume comida rápida	No	Sí
Realiza ejercicio regularmente	Si	No

Tabla 4.3 Antecedentes gineco-obstétricos de casos de estudio

Antecedentes gineco-obstétricos	Caso de estudio 1	Casos de estudio 2, 3 y 4
Duración del ciclo	28 días	35 días
Días del periodo menstrual	5 días	7 días
Edad de inicio de menstruación	13 años	11 años
Edad de inicio de vida sexual	21 años	18 años
Cantidad de embarazos	0	0
Cantidad de abortos	0	0
Tiene alguna enfermedad ginecológica diagnosticada	No	Sí
Lleva tratamiento ginecológico	No	No

Tabla 4.4 Antecedentes familiares en casos de estudio

Antecedentes familiares	Caso de estudio 1	Casos de estudio 2, 3 y 4
Diabetes en familiares	No	Si
Hipertensión en familiares	No	No
SOP en familiares	No	Si

4.1.1 Caso de estudio 1: Sugerencia de asistencia médica en mujer sana.

El primer modelo, como se describió en el capítulo 3, de toda la información que captura la paciente en la aplicación móvil, toma sólo 6 variables: si presenta o no dolor durante las relaciones sexuales, si presenta o no oscuridad en la piel, si presenta o no dolor durante la menstruación, el grupo de edad al que pertenece (calculado a partir de su fecha de nacimiento), si algún familiar padece SOP y la edad de inicio de menstruación. Es importante destacar que el resto de la información es útil para el seguimiento de la salud femenina y, en su caso, se utilizarán en los modelos posteriores, recordando que esta captura previa disminuye el problema de incomodidad en la paciente al asistir al médico y responder preguntas de índole íntima.

Específicamente para el primer caso de prueba, como resultado de los datos descritos en las tablas anteriores, el modelo indica que no es necesario acudir al médico, ya que los valores se encuentran dentro de los parámetros normales. La aplicación muestra un mensaje positivo y recomienda mantener el monitoreo regular de su salud ginecológica (Figura 4.1). Este resultado valida la capacidad del sistema para descartar de forma segura casos que no requieren atención inmediata.

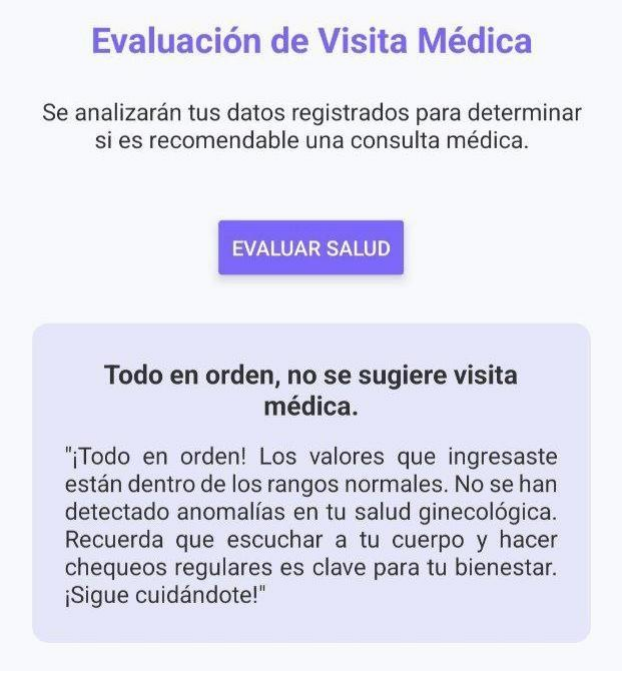

Figura 4.1 Resultado de la Predicción del Primer modelo en paciente sana

4.1.2 Caso de estudio 2: Sugerencia de asistencia médica en mujer con una enfermedad ginecológica

En este caso y, como puede verse en las tablas anteriores, se trata de una persona con antecedentes familiares de SOP, que presenta dolor tanto durante el periodo menstrual como durante las relaciones sexuales y su edad de inicio de menstruación fue relativamente baja. Por lo tanto, tras ejecutar el modelo de clasificación en la aplicación móvil, se obtuvo como resultado un mensaje informativo (Figura 4.2) donde indica que algunos datos están asociados con condiciones que requieren atención médica.

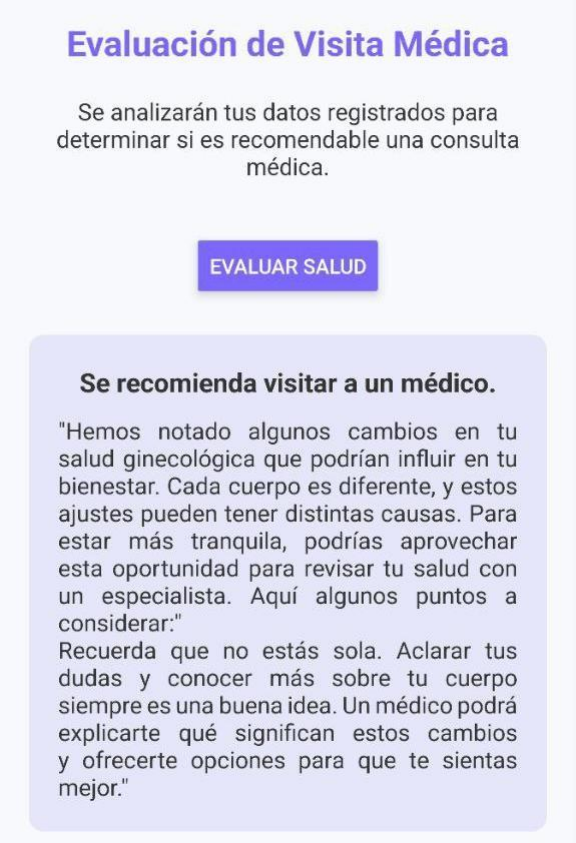

Figura 4.2 Resultado de la Predicción del Primer modelo en paciente con enfermedad ginecológica

4.1.3 Caso de estudio 3: Sugerencia de solicitud de estudios laboratoriales

Tras recibir la sugerencia de buscar asistencia médica, se espera que la paciente agende una cita y comparta sus datos con un profesional de la salud que también utilice la plataforma. Al acudir a la cita, el (la) médico(a) consultará los datos previamente capturados por la paciente y registrará nuevos, en particular los relacionados con hiperandrogenismo (ver el punto 3.1.3 del capítulo anterior).

- Sumatoria de hirsutismo por área del cuerpo: 21

  Donde se consideran las ponderaciones de vello en barbilla, pecho, abdomen superior, abdomen inferior, muslos, brazos, espalda inferior, espalda superior y labio superior.
- Nivel de acné: 3
- Sumatoria de caída del cabello: 12

  Donde se consideran las ponderaciones de frente, mejillas, nariz, barbilla, pecho y espalda superior.

Con todos estos datos, la aplicación web obtiene una clasificación indicando que sí es recomendable solicitar exámenes laboratoriales para obtener un diagnóstico más preciso. Esta sugerencia se muestra con un mensaje claro en la pantalla de resultados, en la cual se especifica el motivo por el que el sistema está realizando esta sugerencia (Figura 4.3), permitiendo al médico justificar y documentar la orden de estudios. Este ejemplo muestra cómo la aplicación web actúa como un sistema de apoyo en la toma de decisiones clínicas.

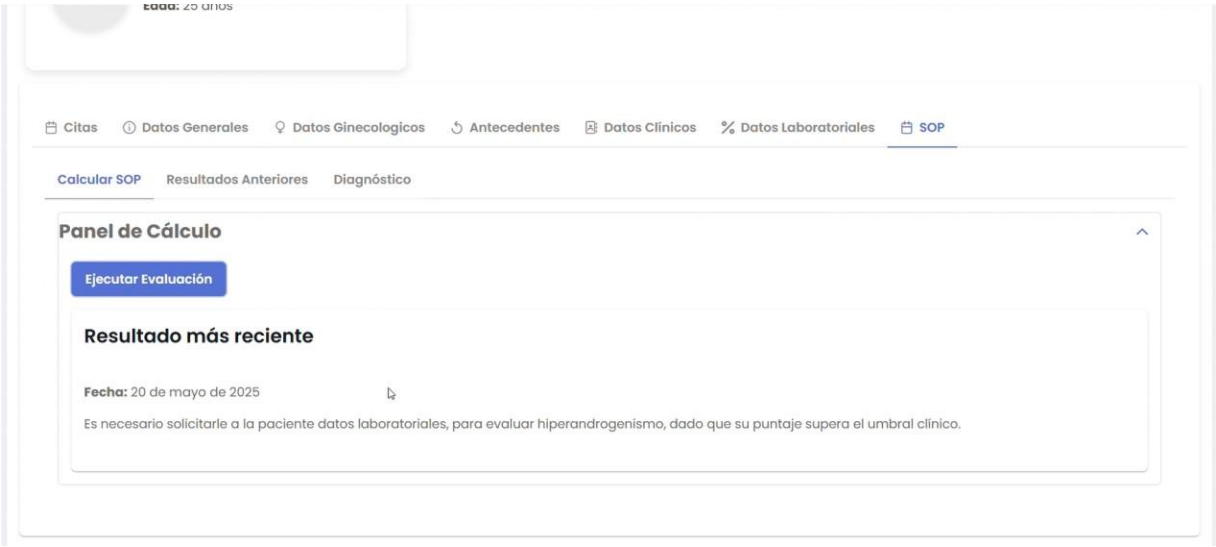

Figura 4.3 Pantalla panel de cálculo de SOP con resultado para solicitar datos laboratoriales

4.1.4 Caso de estudio 4: Sugerencia de continuación de seguimiento con especialista

En este punto se cuenta con la información capturada por la paciente, con los datos de auscultación clínica del (la) médico(a) y los resultados de los estudios laboratoriales, que incluyen Hormona Luteinizante, Hormona Foliculoestimulante, Prolactina, Progesterona entre otros (ver punto 3.1.4 del capítulo anterior).

- Presión arterial sistólica: 130
- Presión arterial diastólica: 88
- Niveles de azúcar en sangre: 125
- FSH (Hormona foliculoestimulante) : 4.8
- LH (Hormona luteinizante): 14
- FSH/LH (Relación Hormona foliculoestimulante / Hormona luteinizante): 0.34
- TSH (hormona estimulante de la tiroides): 3.0
- Prolactina 20
- Progesterona: 1.1

Con la información anterior, la aplicación arroja una clasificación positiva a SOP, indicando que existe una alta probabilidad de que la paciente presente dicha afección, y recomienda continuar el seguimiento con un(a) especialista en Ginecología (Figura 4.4). Este resultado refuerza la utilidad del modelo como herramienta de diagnóstico complementaria, permitiendo al profesional tomar decisiones clínicas basadas en patrones identificados por técnicas de minería de datos.

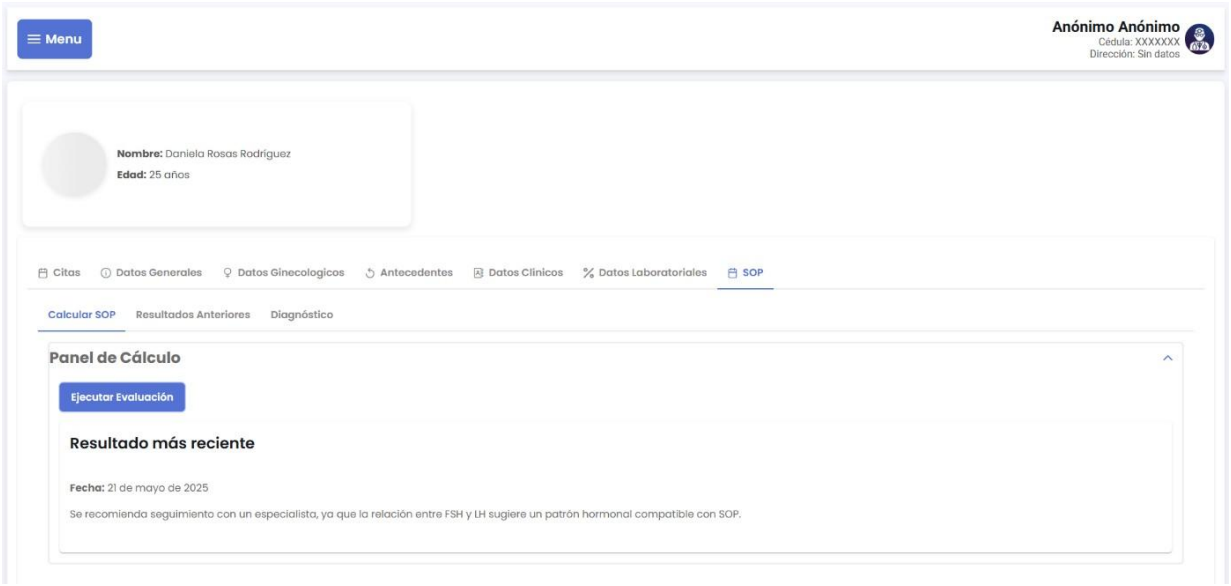

Figura 4.4 Pantalla panel de cálculo de SOP con resultado para sugerir seguimiento con especialista

4.2 Aplicación móvil

La aplicación móvil se diseñó para el uso directo de las pacientes, permitiéndoles registrar de forma accesible sus datos generales y ginecológicos, mediante formularios organizados por secciones; las usuarias tienen la posibilidad de ingresar información referente a sus ciclos menstruales, antecedentes ginecológicos, antecedentes heredofamiliares, entre otros indicadores relevantes para la identificación de una enfermedad ginecológica. Dicha información estará disponible para un médico o especialista con previa autorización de la paciente, esto como medida de privacidad que asegura la confidencialidad de los datos personales de la paciente.

Cada caso ilustra la interacción que tiene la paciente con la interfaz, así como las respuestas generadas por el sistema, mostrando la funcionalidad de los modelos de clasificación y su utilidad en el acompañamiento a la paciente. Para utilizar la aplicación es necesario registrar una nueva cuenta que solicita datos como el correo, contraseñas y nombre, tal como se muestra en la figura 4.5.

Figura 4.5 Pantalla de Registro de Paciente en aplicación móvil

Posteriormente, la usuaria ingresa dichos datos en su inicio de sesión (Figura 4.6). Una vez se encuentre dentro del sistema, se le muestra el menú principal (Figura 4.7), el cual le da acceso a toda la funcionalidad del sistema donde ella ingresa los datos pertinentes relacionados con su salud ginecológica.

Figura 4.6 Pantalla de Inicio de Sesión de paciente en aplicación móvil

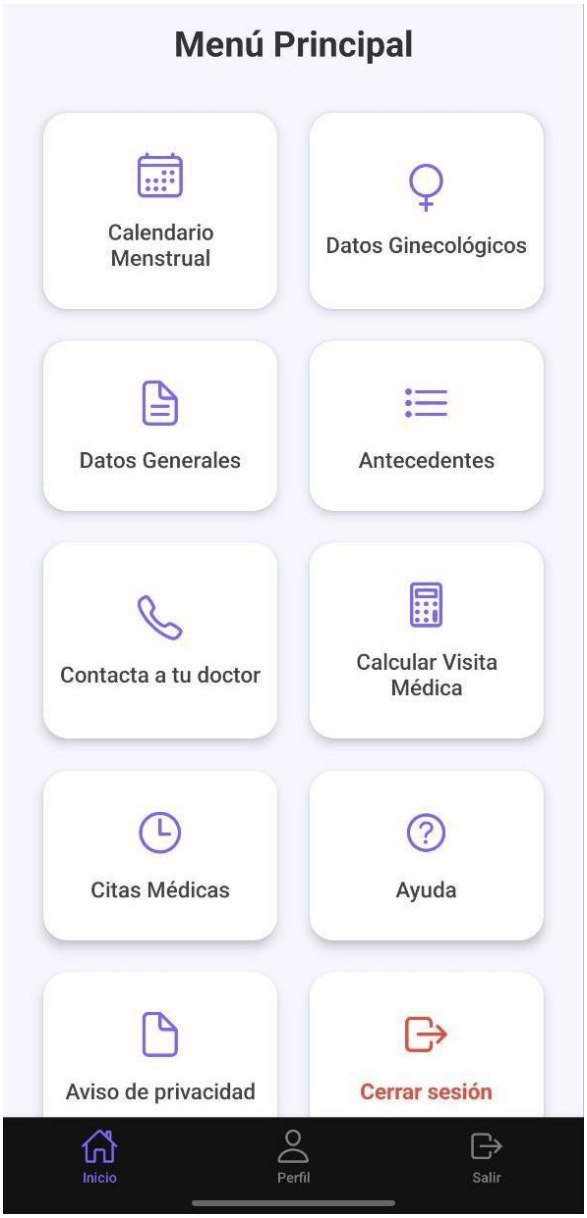

Figura 4.7 Pantalla de Menú Principal de aplicación móvil

Desde la pantalla de inicio de la aplicación, la paciente accede a los tres formularios principales del menú “Datos Ginecológicos” (Figura 4.8), “Datos Generales” (Figura 4.9) y “Datos gineco-obstétricos” (Figura 4.10), completando cada campo solicitado. Estas secciones otorgan al sistema los datos necesarios para ejecutar el primer modelo de clasificación. También se muestran otras opciones como son “Calendario Menstrual”, “Citas médicas” y “Contacta a tu doctor” entre otras; estas secciones permiten llevar un mejor seguimiento de la salud

ginecológica pero no están directamente relacionadas con la identificación de posibles enfermedades. Posteriormente, la usuaria activa el primer modelo de clasificación seleccionando la opción “Calcular Visita Médica” en el menú, diseñado para sugerir a la paciente si es deseable acudir a una consulta médica tomando en cuenta los datos proporcionados.

Es importante indicar que el sistema maneja la información de longitudes en centímetros; sin embargo, por resultar más clara para las pacientes la altura en metros, la app permite la captura en metros, pero envía centímetros al *back-end*.

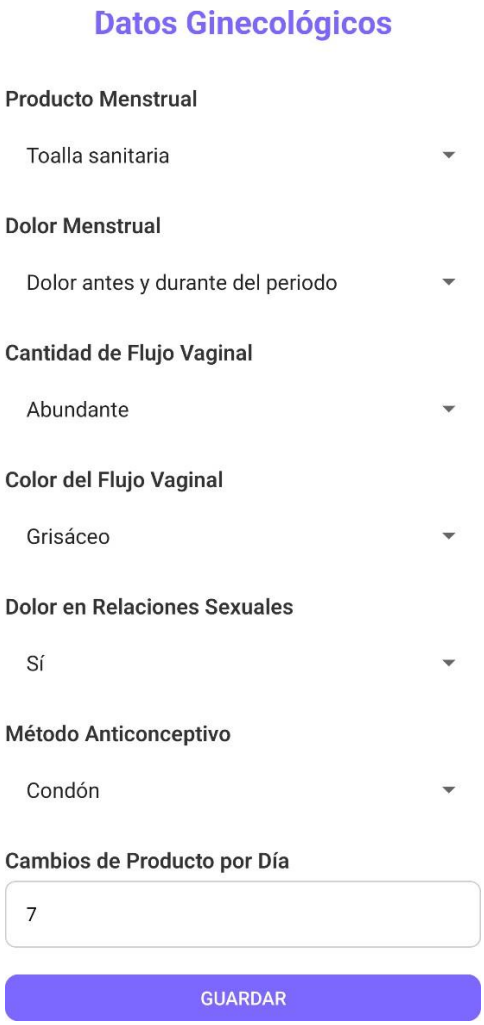

Figura 4.8 Pantalla de Datos Ginecológicos de aplicación móvil

Figura 4.9 Pantalla de Datos generales de aplicación móvil

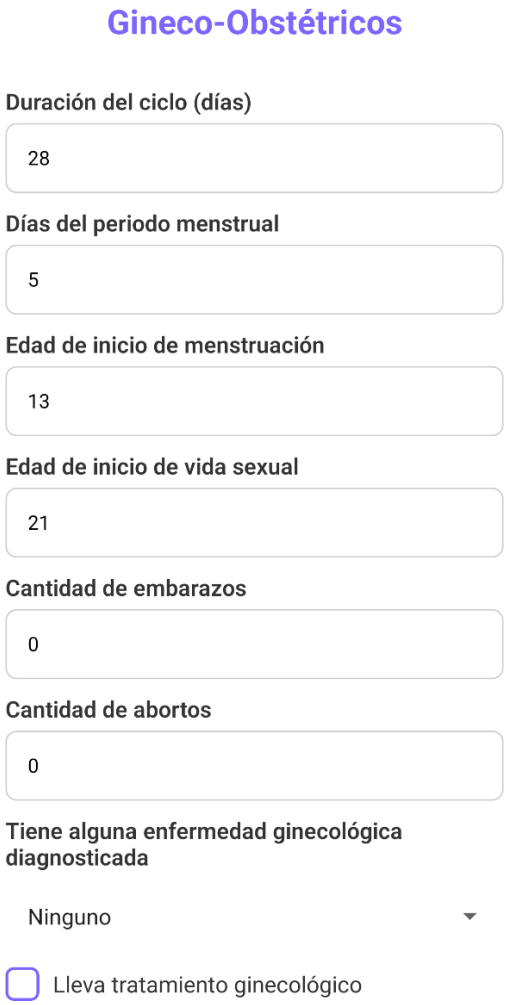

Figura 4.10 Pantalla de Datos gineco-obstétricos de aplicación móvil

4.2 Aplicación web

La aplicación web está orientada a profesionales de la salud, específicamente médicos generales y especialistas en Ginecología, quienes tienen las opciones de acceder a la información de sus pacientes, visualizar las predicciones generadas por los modelos y registrar nuevos datos durante las consultas, esto con el fin de facilitar el seguimiento del estado clínico de cada paciente, permitiendo tomar decisiones informadas sobre la necesidad de estudios laboratoriales o derivaciones médicas con un especialista.

La información de las pacientes se agrega a la agenda del médico a través de tres posibles vías: a) el medico registra directamente a la paciente desde la plataforma incorporándola

manualmente a su agenda, cuando se trata de personas que no tienen dispositivos móviles, no saben cómo utilizar una aplicación para estos dispositivos o no cuentan con fácil acceso a Internet; b) la paciente usa su aplicación móvil y comparte sus datos con el médico, enviándole primero una solicitud que el médico acepta; c) la paciente cuenta con la aplicación para dispositivo móvil pero no envió la solicitud al médico, al llegar a consulta y hablar con el médico identifican que ambos usan la aplicación por lo que el médico envía la solicitud de datos a la paciente en ese momento, la cual será aceptada por la paciente para que concrete la vinculación en la agenda médica.

Cada caso permite mostrar el flujo completo desde la consulta e ingreso de datos hasta la generación de predicciones a partir de ellos. Para acceder al sistema es necesario registrar una nueva cuenta al médico general o especialista, como se ilustra en la figura 4.11 se le solicitan datos como el nombre completo, cédula profesional, foto de perfil, entre otros.

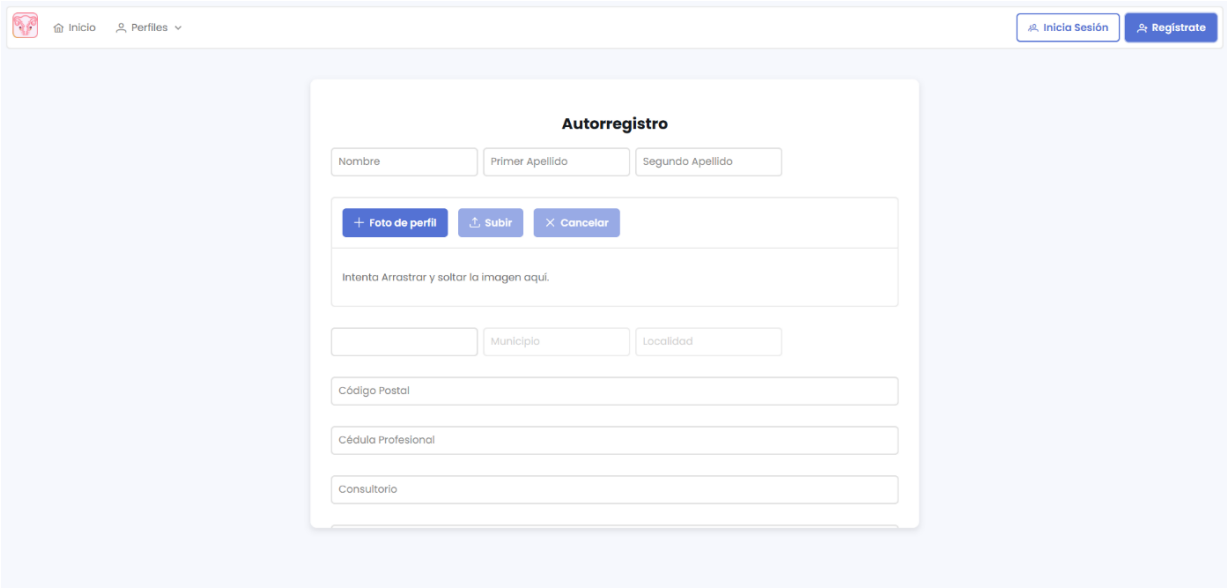

Figura 4.11 Pantalla de Autorregistro médico en aplicación Web

Una vez creada la cuenta, el/la médico ingresa a la plataforma en el inicio de sesión como se muestra en la figura 4.12. Cuando se encuentra dentro del sistema, se le muestra el tablero (*dashboard*) de la aplicación web, los botones y menú para acceder a otras secciones del sistema (Figura 4.13).

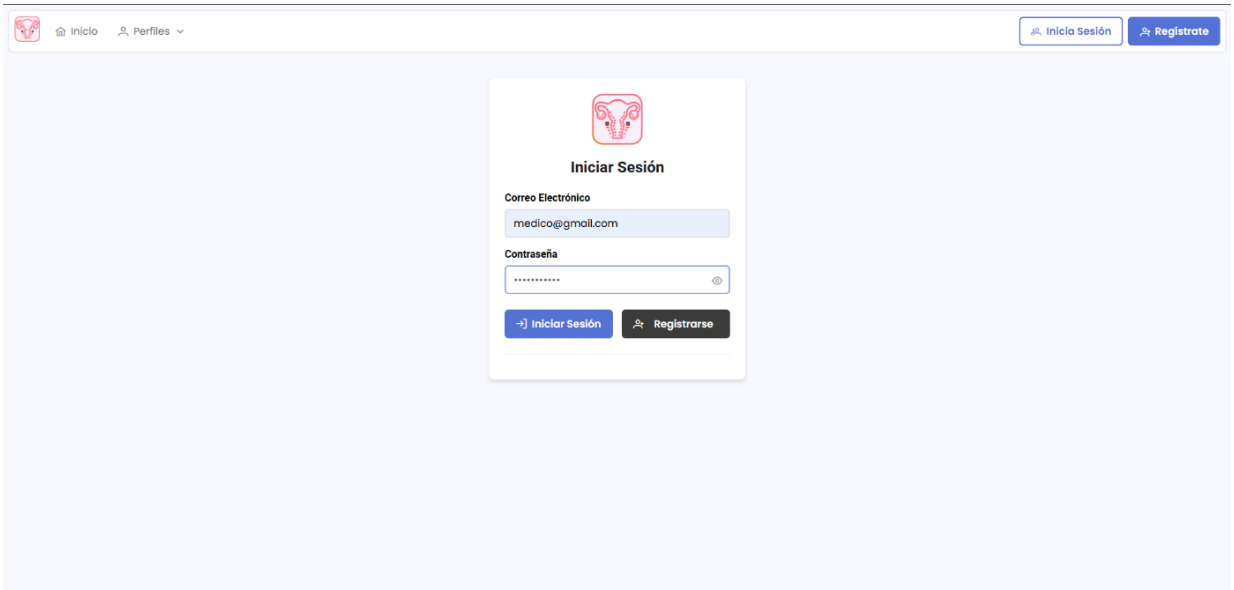

Figura 4.12 Pantalla de Inicio de Sesión de la aplicación Web

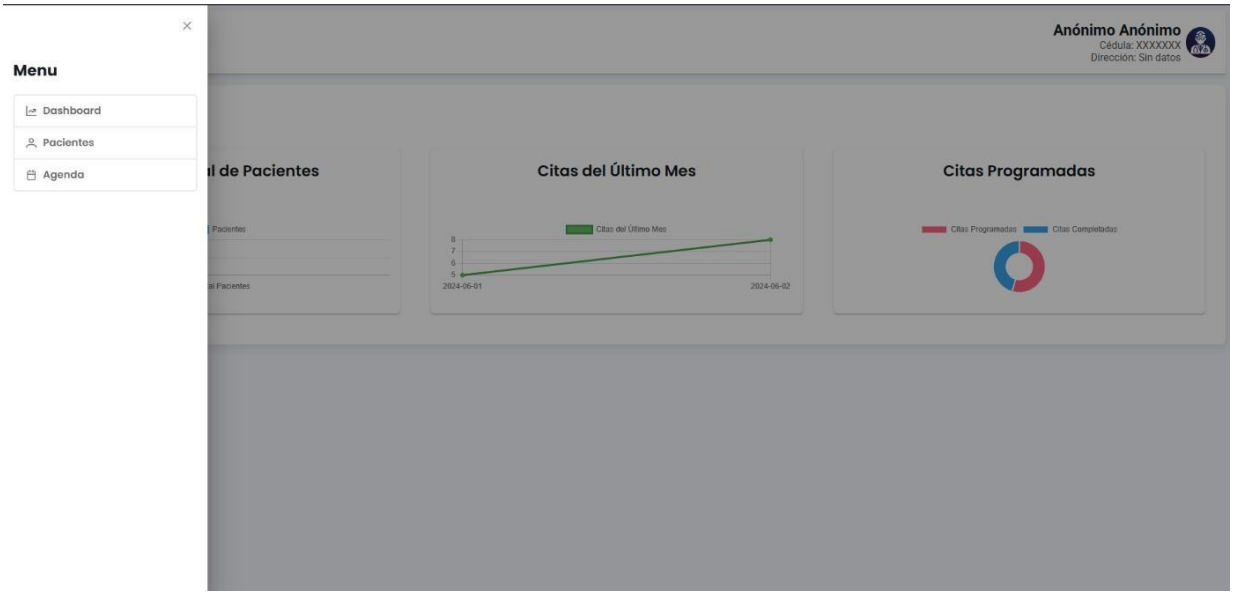

Figura 4.13 Pantalla de *Dashboard* y Menú de aplicación Web

En la sección de “Agenda” (Figura 4.14), a donde se llega mediante una opción del menú, el médico gestionará información de citas y horarios.

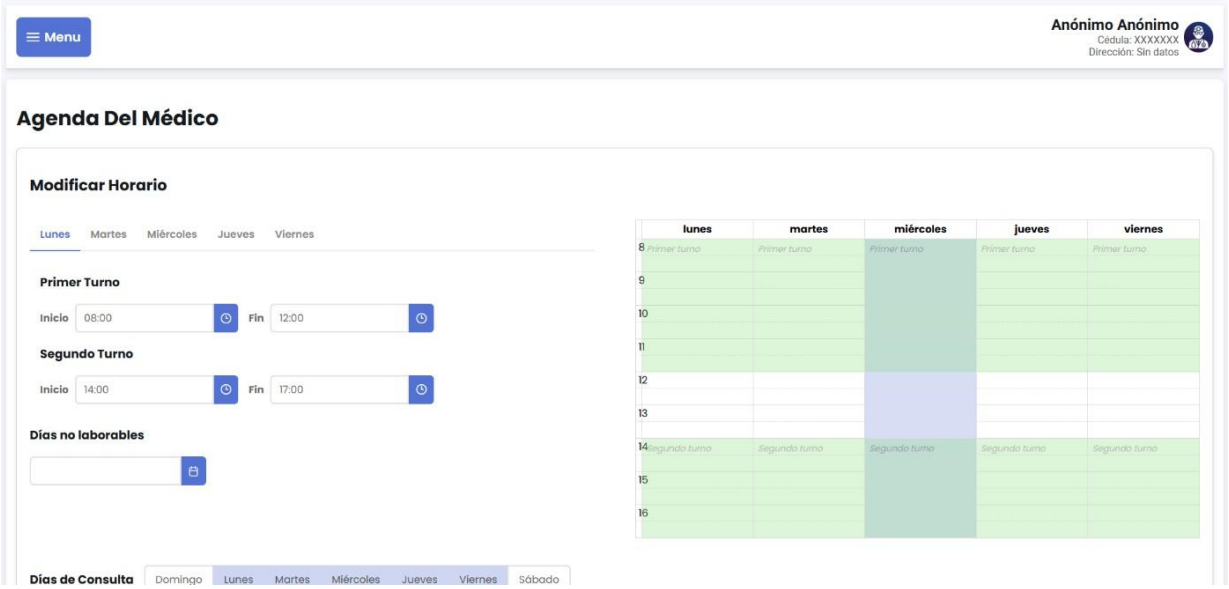

Figura 4.14 Pantalla de Agenda del Médico de aplicación Web

También en el menú se muestra la parte de "Pacientes" (Figura 4.15), donde el personal médico consultará los detalles de sus pacientes y las consultas realizadas.

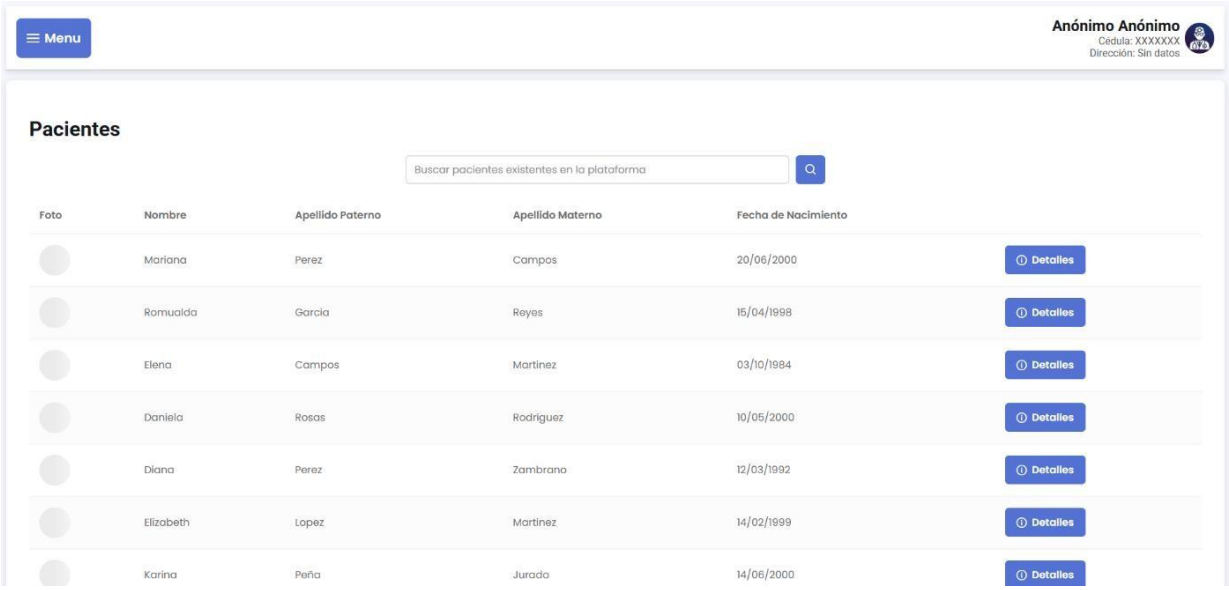

Figura 4.15 Pantalla de Pacientes de aplicación Web

Al momento de la consulta con una paciente, el médico deberá realizar el ingreso en el sistema de los datos clínicos (Figura 4.16), los cuales son necesarios para ejecutar el segundo modelo de clasificación, se incluyen datos como son antecedentes clínicos, hiperandrogenismo y ciclos menstruales, en estos últimos se le muestra al médico una gráfica con los ciclos de los últimos doce meses que ha registrado la paciente desde su aplicación móvil.

Figura 4.16 Pantalla Datos clínicos de aplicación Web

Al acceder a la sección “Datos Laboratoriales” (Figura 4.17) el médico ingresa los valores de laboratorio para su análisis. El tercer modelo de clasificación, diseñado para realizar una clasificación basada en datos laboratoriales, se ejecuta a solicitud del/la médico una vez que se cuenta con estos datos, de otro modo se vuelve a ejecutar el segundo modelo.

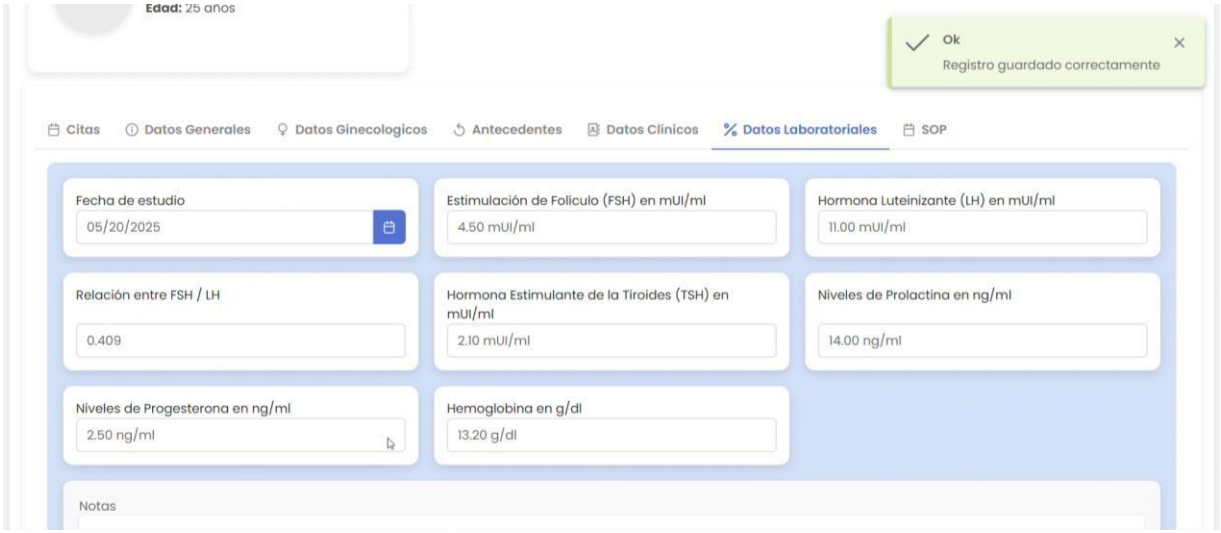

Figura 4.17 Pantalla de Datos Laboratoriales de aplicación Web

Con todo lo anterior se verifica el correcto funcionamiento de ambas aplicaciones y de los modelos de minería de datos implementados en ellas. En el siguiente capítulo se hace énfasis en el cumplimiento de objetivos a partir de los resultados en los casos de prueba aquí mostrados.

Capítulo 5 Conclusiones y recomendaciones

5.1 Conclusiones

La aplicación de técnicas de Inteligencia Artificial en el ámbito de la salud representa una oportunidad significativa para mejorar el diagnóstico temprano y la atención médica oportuna en condiciones complejas como lo es el Síndrome de Ovario Poliquístico (SOP). En este trabajo se diseñó e implementó una plataforma para la salud femenina compuesta por una aplicación para dispositivo móvil y una aplicación web, orientada a un entorno que incluye tanto a pacientes como a profesionales de la salud, con el objetivo de facilitar la detección temprana del SOP y apoyar la toma de decisiones médicas.

Específicamente en la aplicación de técnicas de Minería de Datos, se entrenaron tres modelos de clasificación que actúan en diferentes etapas del proceso de diagnóstico. El primer modelo analiza los síntomas reportados por la paciente y sugiere si es necesario acudir a una consulta médica. El segundo modelo asiste al médico general en la decisión de solicitar estudios laboratoriales y, por último, el tercer modelo permite al médico realizar una predicción basada en datos laboratoriales para determinar el riesgo de efectivamente padecer SOP.

Estos modelos fueron entrenados y validados con datos reales de mujeres en edad reproductiva, logrando niveles de precisión aceptables para su uso en entornos clínicos. Durante las pruebas funcionales, el sistema mostró respuestas positivas con el historial clínico de las pacientes evaluadas, lo cual valida su utilidad como herramienta de apoyo en las decisiones clínicas. La plataforma permite automatizar parte del análisis clínico, reducir el tiempo de obtención de diagnósticos preliminares, aumentar el acceso de las pacientes a información sobre su salud ginecológica y disminuir la reticencia de las pacientes a explicar sus síntomas a un profesional de la salud.

Por otro lado, la integración entre la app y la aplicación web garantiza una interacción eficiente entre paciente y médico, facilitando el seguimiento continuo, pero con un enfoque personalizado a la paciente. El sistema también incluye mensajes explicativos diseñados para informar a las

pacientes sobre su estado de salud ginecológica y en el caso de los médicos estos mensajes están orientados a tener un panorama más amplio con el fin de obtener un diagnóstico más certero. En general, el desarrollo de esta solución demuestra que es posible construir herramientas digitales accesibles, adaptables y orientadas no sólo al médico sino a la paciente, que potencien la atención médica de calidad para mujeres en edad reproductiva, especialmente en contextos donde el acceso a especialistas es limitado.

5.2 Recomendaciones

Uno de los principales retos durante el desarrollo de este trabajo fue la disponibilidad y calidad de los datos ginecológicos, ya que no se encontraron conjuntos de datos con expediente clínico electrónico estandarizado, específicamente para México. Por ello, se recomienda que, en trabajos futuros, se colabore con instituciones que tengan disponibles datos ginecológicos regionales, esto con el fin de facilitar la obtención de conjuntos de datos más amplios, consistentes y actualizados, de manera que los modelos incluyan una regionalización importante.

También sería conveniente ampliar la población de estudio para mejorar el rendimiento de los modelos, especialmente el modelo basado en datos laboratoriales, permitiendo así una validación más robusta y la posibilidad de generar modelos más específicos para diferentes subgrupos de pacientes.

Otra línea de mejora consiste en expandir la funcionalidad del sistema para incluir otras afecciones ginecológicas que utilizan información semejante a la ya considerada. La incorporación de más afecciones femeninas permitiría convertir la plataforma en una herramienta integral de apoyo para la salud ginecológica, aumentando su alcance. Para ello, sería necesario recopilar nuevos conjuntos de datos, adaptar los formularios de ingreso de información en la plataforma y entrenar nuevos algoritmos de clasificación para cada tipo de condición.

Otra mejora posible es independizar la app del *back-end*, permitiendo que sólo requiera acceso a Internet en el momento de compartir los datos con el personal médico, facilitando el uso de la app en lugares con poco o nulo acceso a Internet.

Finalmente, se sugiere continuar con la validación del sistema en entornos clínicos reales, involucrando la opinión de pacientes y más profesionales de la salud, para mejorar tanto la usabilidad como la precisión de las recomendaciones. Este enfoque colaborativo permitirá crecer y mejorar la plataforma para avanzar hacia su integración en centros de atención primaria y ginecológica.

Productos Académicos

Como parte del trabajo de tesis, se realizó el artículo “Análisis de una plataforma para apoyar el diagnóstico del síndrome de ovario poliquístico”, mismo que fue presentado en el CONISOFT 2024 y publicada en la Revista de Investigación en Tecnologías de Información RITI (ISSN 2387-0893)

RITI Journal, Vol. 10, Núm. 27 (Especial 2024) e-ISSN: 2387-0893

RITI SEICIT

Análisis de una plataforma para apoyar en el diagnóstico del síndrome de ovario poliquístico

Analysis of a platform to support the diagnosis of polycystic ovary syndrome

Karen Bozziere Solís
Tecnológico Nacional de México/I.T.Orizaba, Orizaba, Veracruz, México
M18011163@orizaba.tecnm.mx
ORCID: 0009-0003-9079-8748

Beatriz Alejandra Olivares Zepahua
Tecnológico Nacional de México/I.T.Orizaba, Orizaba, Veracruz, México
beatriz.oz@orizaba.tecnm.mx
ORCID: 0000-0003-2799-0887

Laura Nely Sánchez Morales
CONACYT-Tecnológico Nacional de México/I.T.Orizaba, Orizaba, Veracruz, México
laura.sm@orizaba.tecnm.mx
ORCID: 0000-0003-4134-2704

Mónica Ruiz Martínez
Tecnológico Nacional de México/I.T.Orizaba, Orizaba, Veracruz, México
monica.rm@orizaba.tecnm.mx
ORCID: 0009-0002-2379-9524

José Luis Sánchez Cervantes
Tecnológico Nacional de México/I.T.Orizaba, Orizaba, Veracruz, México
jose.sc@orizaba.tecnm.mx
ORCID: 0000-0001-5194-1263

doi: https://doi.org/10.36825/RITI.12.27.008

Recibido: Junio 22, 2024
Aceptado: Agosto 17, 2024

Anexo I - Carta de Intención de Usuario

Tlaltetela, Ver. a 30 de mayo de 2025

ASUNTO: Carta de Usuario

M.C.E. Beatriz Alejandra Olivares Zepahua
Profesor Investigador
Instituto Tecnológico de Orizaba
PRESENTE

Por medio de la presente me dirijo a usted para felicitarle por la terminación del proyecto titulado *"Desarrollo de una plataforma médica para la salud femenina, relacionada con el síndrome de ovario poliquístico, utilizando técnicas de inteligencia artificial"* realizado por la alumna **Karen Bozziere Solís**. Dicho proyecto será de gran utilidad para nosotros ya que nos ayudará en el seguimiento de enfermedades ginecológicas y nos permitirá llegar a un diagnóstico más preciso y rápido. Esta herramienta se implementará en el área de medicina general en el consultorio médico de la comunidad de Tlaltetela, Veracruz, su implementación en este sector facilitará la evaluación integral de casos ginecológicos específicos, así como un mejor seguimiento de las pacientes que visiten el consultorio, permitiendo mejorar las decisiones clínicas.

Atentamente

Dra. Guadalupe Verónica Islas Vargas
Médico General
Cédula 6474743

Referencias

[1] MedlinePlus en español [Internet]. Bethesda (MD): Biblioteca Nacional de Medicina (EE. UU.), "Salud de las mujeres." Accessed: Jul. 30, 2024. [Online]. Available: Medlineplus.gov

[2] F. Concha C., T. Sir P., S. E. Recabarren, and F. Pérez B., "Epigenética del síndrome de ovario poliquístico," *Rev Med Chil*, vol. 145, no. 7, pp. 907–915, Jul. 2017, doi: 10.4067/s0034-98872017000700907.

[3] H. Teede, C. T. Tay, J. S. E. Laven, A. Dokras, L. J. Moran, and T. Piltonen, "International evidence-based guideline for the assessment and management of Polycystic Ovary Syndrome 2023," 2023, *Monash University, Melbourne, Australia*. doi: 10.26180/24003834.V1.

[4] C. M. Francisco, "Alopecia Androgenética Femenina. Etiología y diagnóstico.," *Chilena de Dermatología*, vol. 28, pp. 240–269, 2012.

[5] Pontificia Universidad Católica de Chile, "Hirsutismo: manejo inicial en APS." Accessed: May 28, 2025. [Online]. Available: https://medicina.uc.cl/publicacion/hirsutismo-manejo-inicial-en-aps/

[6] Rotterdam ESHRE/ASRM-Sponsored PCOS consensus workshop group, "Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome (PCOS)," *Human Reproduction*, vol. 19, no. 1, Jan. 2004, doi: 10.1093/humrep/deh098.

[7] H. J. Teede *et al.*, "Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome," *Fertil Steril*, vol. 110, no. 3, pp. 364–379, Aug. 2018, doi: 10.1016/j.fertnstert.2018.05.004.

[8] P. Hamet and J. Tremblay, "Artificial intelligence in medicine," *Metabolism*, vol. 69, pp. S36–S40, Apr. 2017, doi: 10.1016/j.metabol.2017.01.011.

[9] E. Buulolo, S.Kom, and M.Kom, *Data Mining Untuk Perguruan Tinggi*. Deepublish, 2020.

[10] V. DEVEDZIC, “KNOWLEDGE DISCOVERY AND DATA MINING IN DATABASES,” 2001, pp. 615–637. doi: 10.1142/9789812389718_0025.

[11] Data Science Process Alliance, “KDD and Data Mining.” Accessed: May 29, 2025. [Online]. Available: https://www.datascience-pm.com/kdd-and-data-mining/

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[13] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*, 3rd ed. United States of America: Wiley-IEEE Press., 2019.

[14] IBM, “Naive Bayes.” Accessed: May 29, 2025. [Online]. Available: https://www.ibm.com/think/topics/naive-bayes

[15] K. Y. Chan *et al.*, “Deep neural networks in the cloud: Review, applications, challenges and research directions,” *Neurocomputing*, vol. 545, p. 126327, Aug. 2023, doi: 10.1016/j.neucom.2023.126327.

[16] A. Chellam, R. L, and R. S, “Intrusion Detection in Computer Networks using Lazy Learning Algorithm,” *Procedia Comput Sci*, vol. 132, pp. 928–936, 2018, doi: 10.1016/j.procs.2018.05.108.

[17] W. Ian H, F. Eibe, H. Mark A, and P. Christopher J, *Data mining: practical machine learning tolos and techniques*, 4th ed. 2016.

[18] scikit-learn developers, “RandomForestClassifier,” Documentación oficial de scikit-learn. Accessed: May 07, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[19] scikit-learn developers, “sklearn.metrics.accuracy_score”, Accessed: May 09, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

[20] scikit-learn developers, “sklearn.metrics.cohen_kappa_score,” Documentación oficial de scikit-learn. Accessed: May 07, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

[21] Amazon Web Services, “Metrics and validation – Amazon SageMaker.” Accessed: May 07, 2025. [Online]. Available: https://docs.aws.amazon.com/sagemaker/latest/dg/autopilot-metrics-validation.html

[22] E. Zhang and Y. Zhang, “F-Measure,” in *Encyclopedia of Database Systems*, Boston, MA: Springer US, 2009, pp. 1147–1147. doi: 10.1007/978-0-387-39940-9_483.

[23] A. Piyush, “Feature Selection Techniques in Machine Learning,” Kaggle. Accessed: May 30, 2025. [Online]. Available: https://www.kaggle.com/code/piyushagni5/feature-selection-techniques-in-machine-learning

[24] Weka Documentation, “weka.attributeSelection.CfsSubsetEval,” Weka Documentation. Accessed: May 29, 2025. [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/attributeSelection/CfsSubsetEval.html

[25] Weka Documentation, “weka.attributeSelection.BestFirst.” Accessed: May 29, 2025. [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/attributeSelection/BestFirst.html

[26] D. Moloy, “Chi-Square and Machine Learning,” IBM Community Blog. Accessed: May 29, 2025. [Online]. Available: https://community.ibm.com/community/user/blogs/moloy-de1/2023/06/12/chi-square-and-machine-learning

[27] Weka Documentation, “weka.attributeSelection.InfoGainAttributeEval.” Accessed: May 30, 2025. [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html

[28] Weka Documentation, “weka.attributeSelection.ReliefFAttributeEval.” Accessed: May 30, 2025. [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/attributeSelection/ReliefFAttributeEval.html

[29] Amazon Web Services (AWS), “¿Qué es Scrum?,” Amazon Web Services (AWS). Accessed: May 29, 2025. [Online]. Available: https://aws.amazon.com/es/what-is/scrum/

[30] Cognizant, “Plataforma digital.” Accessed: Jul. 30, 2024. [Online]. Available: https://www.cognizant.com/es/es/glossary/digital-platform

[31] GCFGlobal, “Informática Básica: ¿Qué es una aplicación móvil?” Accessed: Jul. 30, 2024. [Online]. Available: https://edu.gcfglobal.org/es/informatica-basica/que-es-una-aplicacion-movil/1/

[32] Secretaria de salud, *NORMA Oficial Mexicana, Del expediente clínico*. México, 2010.

[33] Cámara de Diputados, “Ley General de Protección de Datos Personales en Posesión de Sujetos Obligados.” Accessed: Aug. 18, 2024. [Online]. Available: http://www.diputados.gob.mx/LeyesBiblio/ref/lgpdppso.htm

[34] MedlinePlus en español [Internet]. Bethesda (MD): Biblioteca Nacional de Medicina (EE. UU.), “Pruebas de laboratorio.” Accessed: Aug. 18, 2024. [Online]. Available: https://medlineplus.gov/spanish/laboratorytests.html

[35] Unidad de Planeación y Evaluación de Programas para el Desarrollo, “Informe anual sobre la situación de pobreza y rezago social 2022,” Tlaltetela, 2022. Accessed: Aug. 18, 2024. [Online]. Available: https://www.gob.mx/cms/uploads/attachment/file/698603/30_024_VER_Tlaltetela.pdf

[36] Consejo Nacional de Población, “Día Internacional de Acción por la Salud de las Mujeres.” Accessed: Jul. 30, 2024. [Online]. Available: https://www.gob.mx/conapo/articulos/dia-internacional-de-accion-por-la-salud-de-las-mujeres-303826?idiom=es.

[37] T. Corona Vázquez, M. E. Medina Mora, P. Ostrosky Wegman, E. J. Sarti Gutiérrez, and P. Uribe Zúñiga, *La mujer y la salud en México*. Intersistemas S.A de C.V, 2014.

[38] Mayo Clinic, “Síndrome de ovario poliquístico.” Accessed: Jul. 30, 2024. [Online]. Available: https://www.mayoclinic.org/es/diseases-conditions/pcos/symptoms-causes/syc-20353439.

[39] P. S. Vanhauwaert, “Síndrome de ovario poliquístico e infertilidad,” *Revista Médica Clínica Las Condes*, vol. 32, no. 2, pp. 166–172, Mar. 2021, doi: 10.1016/j.rmclc.2020.11.005.

[40] Food and Travel, “Apps para el cuidado femenino.” Accessed: Jul. 30, 2024. [Online]. Available: https://foodandtravel.mx/apps-para-el-cuidado-femenino/.

[41] Instituto Nacional de Estadística y Geografía, “Encuesta Nacional de la Dinámica Demográfica (ENADID) 2018.” Accessed: Aug. 08, 2024. [Online]. Available: https://www.inegi.org.mx/programas/enadid/2018/

[42] Cámara de Diputados, “Entre 6 y 10 por ciento de las mexicanas padece Síndrome del Ovario Poliquístico.” Accessed: Aug. 01, 2024. [Online]. Available: http://www5.diputados.gob.mx/index.php/esl/Comunicacion/Boletines/2017/Julio/31/3888-Entre-6-y-10-por-ciento-de-las-mexicanas-padece-Sindrome-del-Ovario-Poliquistico

[43] S. Bharati, P. Podder, and M. R. Hossain Mondal, “Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms,” in *2020 IEEE Region 10 Symposium (TENSYMP)*, IEEE, 2020, pp. 1486–1489. doi: 10.1109/TENSYMP50017.2020.9230932.

[44] M. Soomro and A. Akhlaq, “Impact of Cyberchondriasis on Polycystic Ovarian Syndrome Patients Searching Health Information Online,” in *Proceedings of the 2018 International Conference on Digital Health*, New York, NY, USA: ACM, Apr. 2018, pp. 60–64. doi: 10.1145/3194658.3194685.

[45] S. Ramamoorthy, V. R., and R. Sivasubramaniam, “Monitoring the growth of Polycystic Ovary Syndrome using Mono-modal Image Registration Technique,” in *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, New York, NY, USA: ACM, Jan. 2019, pp. 180–187. doi: 10.1145/3297001.3297024.

[46] M. A. Erandathi, W. Y. C. Wang, and M. Mayo, “Predicting Diabetes Mellitus and its Complications through a Graph-Based Risk Scoring System,” in *Proceedings of the 4th International Conference on Medical and Health Informatics*, New York, NY, USA: ACM, Aug. 2020, pp. 1–7. doi: 10.1145/3418094.3418115.

[47] R. Deo and S. Panigrahi, “Prediction of Hepatic Steatosis (Fatty Liver) using Machine Learning,” in *Proceedings of the 2019 3rd International Conference on Computational Biology and Bioinformatics*, New York, NY, USA: ACM, Oct. 2019, pp. 8–12. doi: 10.1145/3365966.3365968.

[48] P. Wang and W. Zhu, “Reproductive history and breast cancer risk in women: based on logistic regression and survival analysis,” in *2022 4th International Conference on*

Intelligent Medicine and Image Processing, New York, NY, USA: ACM, Mar. 2022, pp. 1–7. doi: 10.1145/3524086.3524087.

[49] S. E. Fox, A. Menking, J. Eschler, and U. Backonja, "Multiples Over Models," *ACM Transactions on Computer-Human Interaction*, vol. 27, no. 4, pp. 1–24, Aug. 2020, doi: 10.1145/3397178.

[50] S. Chopra, R. Zehrung, T. A. Shanmugam, and E. K. Choe, "Living with Uncertainty and Stigma: Self-Experimentation and Support-Seeking around Polycystic Ovary Syndrome," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, May 2021, pp. 1–18. doi: 10.1145/3411764.3445706.

[51] C. M. Wu, M. Badshah, and V. Bhagwat, "Heart Disease Prediction Using Data Mining Techniques," in *Proceedings of the 2019 2nd International Conference on Data Science and Information Technology*, New York, NY, USA: ACM, Jul. 2019, pp. 7–11. doi: 10.1145/3352411.3352413.

[52] Y. Li and D. Li, "University Students' Behavior Characteristics Analysis and Prediction Method Based on Combined Data Mining Model," in *Proceedings of the 2020 the 3rd International Conference on Computers in Management and Business*, New York, NY, USA: ACM, Jan. 2020, pp. 9–13. doi: 10.1145/3383845.3383868.

[53] I. Jahangir, Abdul-Basit, A. Hannan, and S. Javed, "Prediction of Dengue Disease through Data Mining by using Modified Apriori Algorithm," in *Proceedings of the 4th ACM International Conference of Computing for Engineering and Sciences*, New York, NY, USA: ACM, Jul. 2018, pp. 1–4. doi: 10.1145/3213187.3287612.

[54] H. Elmannai *et al.*, "Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and Explainable Artificial Intelligence," *Diagnostics*, vol. 13, no. 8, p. 1506, Apr. 2023, doi: 10.3390/diagnostics13081506.

[55] S. Chang and A. Dunaif, "Diagnosis of Polycystic Ovary Syndrome," *Endocrinol Metab Clin North Am*, vol. 50, no. 1, pp. 11–23, Mar. 2021, doi: 10.1016/j.ecl.2020.10.002.

[56] A. E. Joham, T. Piltonen, M. E. Lujan, S. Kiconco, and C. T. Tay, "Challenges in diagnosis and understanding of natural history of polycystic ovary syndrome," *Clin Endocrinol (Oxf)*, vol. 97, no. 2, pp. 165–173, Aug. 2022, doi: 10.1111/cen.14757.

[57] S. A. Suha and M. N. Islam, “An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image,” *Sci Rep*, vol. 12, no. 1, p. 17123, Oct. 2022, doi: 10.1038/s41598-022-21724-0.