MÁSTER OFICIAL INTERUNIVERSITARIO EN INGENIERÍA BIOMÉDICA

TRABAJO FIN DE MÁSTER
Trabajo de investigación

Desarrollo de modelos predictivos y una aplicación móvil para la predicción de la depresión postparto

Santiago Jiménez Serrano
sanjiser@upv.es

Dirigido por:
Dr. Salvador Tortajada Velert
vesaltor@upv.es

Dr. Juan Miguel García-Gómez
juanmig@ibime.upv.es

Valencia - Septiembre 2013

Resumen

La depresión postparto es una enfermedad universal incapacitante que afecta a muchas mujeres de diferentes países durante una etapa vital clave para ella y su recién nacido. El objetivo del presente trabajo ha consistido en la creación de modelos de clasificación de la depresión postparto basados en datos clínicos y cuestionarios a pacientes. Se analizaron dichos datos y se aplicó una metodología de experimentación para el desarrollo, validación y evaluación de diferentes modelos de clasificación. Los clasificadores que presentaron un mejor balance entre sensibilidad y especificidad se integraron en un sistema de ayuda a la decisión clínica para plataformas móviles Android. Se ha pretendido pues, poner en manos tanto de personal clínico como de pacientes una herramienta que ayude en la prevención de la enfermedad y en la detección de población de riesgo. Es por eso que la aplicación final se presenta en dos versiones, una para expertos médicos, y otra más sencilla para las mujeres que acaban de dar a luz.

Palabras Clave

Depresión postparto, redes neuronales, Android, clasificadores, ayuda a la decisión, embarazo.

TREBALL FINAL DE MÀSTER
Treball de recerca

Títol:	Desenvolupament de models predictius i una aplicació mòbil per a la predicció de la depressió postpart
Titulació:	Màster Oficial Interuniversitari en Enginyeria Biomèdica
Autor:	Santiago Jiménez Serrano <sanjiser@upv.es>
Directors:	Dr. Salvador Tortajada Velert <vesaltor@upv.es> Dr. Juan Miguel García-Gómez <juanmig@ibime.upv.es>
Data:	Setembre de 2013

Resum

La depressió postpart es una malaltia universal incapacitant que afecta a moltes dones de diferents països durant una etapa vital clau per a ella i el seu nadó. L'objectiu del present treball ha estat la creació de models de classificació de la depressió postpart basats en dades clíniques i qüestionaris a pacients. Es van analitzar aquestes dades i es va aplicar una metodologia d'experimentació per al desenvolupament, validació i avaluació de diferents models de classificació. Els classificadors que presentaren un millor balanç entre sensibilitat i especificitat s'integraren en un sistema d'ajuda a la decisió clínica per a plataformes mòbils Android. S'ha pretès, doncs, posar a les mans de personal clínic i de pacients una ferramenta que ajude en la prevenció de la malaltia i en la detecció de població de risc. És per això que l'aplicació final es presenta en dues versions, una per a experts mèdics, i una altra més senzilla per a les dones que acaben de donar a llum.

Paraules Clau

Depressió postpart, xarxes neuronals, Android, classificadors, ajuda a la decisió, embaràs.

FINAL MASTER'S PROJECT
Research work

Title:	Development of predictive models and a mobile application for the prediction of postpartum depression
Degree:	Master's Degree in Biomedical Engineering
Author:	Santiago Jiménez Serrano <sanjiser@upv.es>
Directors:	Dr. Salvador Tortajada Velert <vesaltor@upv.es> Dr. Juan Miguel García-Gómez <juanmig@ibime.upv.es>
Date:	September 2013

Abstract

Postpartum depression is a disabling universal disease that affects many women from different countries during a key life stage for her and her newborn. The aim of this research involved the creation of classification models for postpartum depression based on clinical and patient questionnaires. We analyzed these data and applied a methodology of experimentation, validation and evaluation of different classification models. The classifiers with the best balance between sensitivity and specificity were integrated into a clinical decision support system for Android mobile platforms. It is intended, therefore, to put in clinicians and patients hands, a tool that aims to prevent the disease as well as to detect the risk population. That's why the final application comes in two versions, one for medical experts, and another one simpler for women who have just given birth.

Keywords

Postpartum depression, neural networks, Android, classifiers, decision support, pregnancy.

Índice de Contenidos

Índice de Figuras

Índice de Tablas

Acrónimos

AA	Aprendizaje Automático
ACC	*Accuracy* - Precisión
ANN	*Artificial Neural Networks* - Redes Neuronales Artificiales
AUC	*Area Under the Curve* - Área bajo la curva
BDI	*Beck Depression Inventory*
DIGS	*Diagnostic Interview for Genetic Studies*
DPP	Depresión Postparto
DSM-IV-TR	Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision
EPDS	*Edinburgh Postnatal Depression Scale* – Escala de depresión postparto de Edimburgo
EPQ	*Eysenck Personality Questionnaire*
ESP	Especificidad
FP	Falsos Positivos
FN	Falsos Negativos
IC	Intervalo de Confianza
ICD	*International Classification of Diseases* – Clasificación internacional de enfermedades
IGU	Interfaz Gráfica de Usuario
KNN	*K-Nearest Neighbors* – K-Vecinos más cercanos
OMS	Organización Mundial de la Salud
PPD	*Postpartum Depression*
RNA	Redes Neuronales Artificiales
RDC	*Research Diagnostic Criteria*
ROC	*Receiver Operating Characteristic*
RP	Reconocimiento de Patrones
RV+	Razón de Verosimilitud Positiva
RV-	Razón de Verosimilitud Negativa
SADM	Sistemas de Ayuda a la Decisión Médica
SEN	Sensibilidad
SVM	*Support Vector Machines*
VN	Verdaderos Negativos
VP	Verdaderos Positivos
VPN	Valor Predictivo Negativo
VPP	Valor Predictivo Positivo
ZDI	*Zung Depression Inventory*

1. Introducción

La depresión postparto (DPP) aparece en la 'International Classification of Diseases' ICD-10 de la Organización Mundial de la Salud (OMS) en la categoría F53, relativa a trastornos mentales y del comportamiento asociados con el puerperio, no clasificados en otro lugar [1].

En el DSM-IV-TR (Diagnostic and Statistical Manual of Mental Disorders [2]) se reconoce como un trastorno depresivo mayor de inicio en el postparto, entendiéndose por postparto las primeras cuatro a seis semanas siguientes al mismo. Este intervalo deriva de los datos de ingresos hospitalarios en puérpera por enfermedad psiquiátrica grave y no responde a una razón fisiopatológica clara. Por estos motivos y por las dificultades diagnósticas, la mayoría de los autores consideran un período de inicio más amplio para hablar de depresión postparto, pudiendo incluso llegar hasta pasados los tres o seis meses [3]. Existen autores que alargan esta posibilidad de inicio para desarrollar una DPP hasta un año después del nacimiento, siendo lo normal que aparezca durante las primeras semanas [4]. La duración de la DPP está documentada desde los tres hasta los catorce meses y los datos sugieren que dicha duración está relacionada con la severidad de la enfermedad [5].

El puerperio es un período durante el cual hay una adaptación entre el hijo y la madre. Los sentimientos de ansiedad, irritación, tristeza e inquietud son comunes en las dos primeras semanas después del embarazo, denominándose 'tristeza puerperal' o *'baby blues'* y suelen desaparecer pronto sin necesidad de tratamiento. La DPP puede ocurrir cuando la tristeza puerperal no desaparece, o si lo hiciera, también puede aparecer incluso unos meses después del parto tal como se ha mencionado anteriormente. Sus síntomas son los mismos que una depresión mayor en otros momentos de la vida. El cuadro clínico está caracterizado por la presencia de tristeza, pérdida de la capacidad para experimentar placer, cambios en el sueño y en el apetito, cansancio físico, falta de concentración, e ideas de minusvalía, desesperanza y culpa relacionadas comúnmente con el cuidado y la crianza de los hijos. Las alteraciones en la concentración y las ideas depresivas constituyen la dimensión cognitiva de la enfermedad y suelen ser los síntomas más específicos y sugestivos de la presencia de DPP, al igual que las ideas de muerte, suicidio, y de agresividad hacia el recién nacido [6]. A diferencia de lo que ocurre con los trastornos depresivos en otras etapas de la vida de la mujer, la DPP genera un doble impacto negativo, ya que afecta tanto a la madre, quien padece directamente la enfermedad, como al recién nacido, ya que depende de los cuidados de ésta para su bienestar físico y emocional [7]. Una madre con DPP puede ser incapaz de cuidar de sí misma o de su bebé, o sentir temor a quedarse sola a su cargo. También pudiera tener sentimientos negativos hacia el niño, llegando a pensar en hacerle daño, o preocuparse en exceso o en defecto por él [8]. Esto podría ocasionar un impacto negativo en el futuro desarrollo del infante hasta la escuela primaria [9]. Las investigaciones han demostrado consistentemente que la depresión o ansiedad durante el embarazo, los eventos vitales estresantes y recientes, la falta de apoyo social, la baja autoestima o el neuroticismo son factores de riesgo [10].

Respecto a la epidemiología, existe gran divergencia de números dependiendo de los estudios consultados. Algunos afirman que entre un 8 y un 25% de las mujeres presentan un síndrome depresivo en los meses siguientes al parto [11]. Otros hablan de una prevalencia equivalente en diferentes países de alrededor del 13% [12], y artículos más recientes mencionan un rango entre el 10 y el 18% [13]. Esta variación en los índices de prevalencia responde a la heterogeneidad en los diseños de las investigaciones. Los factores que influyen en la variabilidad de la tasa de prevalencia son el método de selección, el tamaño de la muestra, la definición del período posparto y el método de evaluación utilizado para determinar la depresión, entre otros. Los estudios que utilizan sistemas de diagnóstico estandarizados para identificar la depresión como el DSM, el *Research Diagnostic Criteria* (RDC), o el *Goldberg Criteria* obtienen tasas de prevalencia próximas al 12%. Sin embargo los estudios que emplean medidas de autoinforme como el *Edinburgh Postnatal Depression Scale* (EPDS), *Beck Depression Inventory* (BDI) o *Zung Depression Inventory* (ZDI) encuentran tasas próximas al 14% [14]. Entre las medidas de autoinforme, la más conocida y utilizada para detectar posibles casos de DPP es la EPDS [15] [16], siendo una de las variables que utilizaremos más adelante. Esta escala ha sido traducida a más de 10 idiomas y validada en diferentes países, entre ellos España [11]. Si nos atenemos a la revisión bibliográfica a la que hace referencia la OMS, del sistema público de salud de Toronto (Canadá), la DPP es un problema de salud pública que afecta aproximadamente al 13% de las mujeres dentro del primer año tras dar a luz [10].

Queda evidenciado que la DPP es un trastorno grave que provoca gran sufrimiento tanto a la madre como a la familia, deteriorando su calidad de vida y afectando a la salud del recién nacido [17] [18] [19]. A pesar de ello, muchas veces esta enfermedad no es diagnosticada. En países desarrollados como Estados Unidos cerca del 50% de los casos de DPP continúan sin ser detectados en la práctica clínica [20]. Aunque se han tomado algunas medidas para detectar la sintomatología depresiva en las mujeres que acaban de dar a luz, el desarrollo de un programa de cribado requiere de un esfuerzo considerable y cuidadoso. Las decisiones basadas en la evidencia tienen que tomarse de acuerdo a los dos principios siguientes. El primero sería obtener el test de cribado más efectivo que no sólo tenga una buena sensibilidad y especificidad, sino que además sea rápido y fácil de interpretar y de llevar a la práctica, además de ser sensible culturalmente. Y el segundo principio a considerar es que se han de tener en cuenta las cuestiones relacionadas con un sistema público de salud, tales como el coste-efectividad o los daños potenciales de un mal diagnóstico.

Durante los últimos años existe un especial interés en el diagnóstico y tratamiento tempranos con el fin de poder iniciar el tratamiento integral, y así evitar o limitar las posibles secuelas. Una predicción temprana podría reducir el impacto de la enfermedad en la madre, y ayudaría a los clínicos a dar un tratamiento apropiado a las pacientes para prevenir la depresión. Como consecuencia, ha surgido la necesidad de diseñar instrumentos y herramientas para la detección precoz que puedan ser eficaces y de fácil aplicación por médicos generales, ginecólogos y pediatras, quienes son los profesionales que más frecuentemente tienen contacto con estas pacientes en el periodo gestacional y puerperal.

Es aquí donde es muy interesante considerar el uso de Sistemas de Ayuda a la Decisión Médica (SADM). Estos sistemas informáticos proveen conocimiento específico y preciso para la toma de decisiones que se han de adoptar durante diagnósticos, pronósticos, tratamientos y gestión de pacientes. Los SADM están totalmente relacionados al concepto de medicina basada en la evidencia [21], ya que infieren conocimiento a partir de los datos extraídos de bases de datos biomédicas, para posteriormente ayudar en el diagnóstico de nuevos pacientes en base a dicho conocimiento adquirido.

Los SADM más complejos están directamente basados en el uso de la inteligencia artificial en medicina [22], en concreto con la rama de Aprendizaje Automático (AA), más conocida por su denominación en inglés como *Machine Learning* [23]. En ella se utilizan algoritmos para inferir modelos predictivos a partir de datos del 'mundo real'. En AA, el Reconocimiento de Patrones (RP) se refiere a la asignación de una etiqueta o clase a un valor de entrada. Un ejemplo de RP son los problemas de clasificación, en los que se intenta asignar a cada valor de entrada en una de las clases de un conjunto dado. En el caso que nos ocupa, el ejemplo más directo sería determinar si una madre que acaba de dar a luz tiene o no riesgo de padecer DPP.

Tras una revisión bibliográfica en busca de investigaciones o trabajos que aborden el problema de la predicción temprana de la DPP, encontramos que la mayoría utilizan la entrevista clínica basada en el DSM y los cuestionarios EPDS y BDI para su diagnóstico durante las primeras semanas después del parto [24] [25] [26] [27].

También encontramos que durante 2013, en un pequeño estudio con 51 mujeres embarazadas, investigadores de la *'Johns Hopkins University School of Medicine'*, hallaron cambios en dos genes (HP1BP3 y TTC9B) que podían distinguir entre mujeres en riesgo de sufrir DPP y las que no [13]. Estos cambios son descritos como modificaciones epigenéticas que reaccionan ante cambios en los niveles de estrógenos, y pueden ser detectados a través de un análisis de sangre. Los resultados que obtuvieron fueron de un 85% de precisión y una AUC de 0.96 en el mejor de los casos. Sin embargo, en sus conclusiones hacen notar el pequeño tamaño de la muestra y la necesidad de efectuar un estudio más grande.

Puesto que no existe una única vía etiológica mediante la cual aparece la DPP, es improbable que una única modalidad de prevención o tratamiento sea eficaz para todas las mujeres. Un enfoque multifactorial que combine las contribuciones de los factores psicológicos, psicosociales y biológicos, probablemente sea más beneficioso y eficaz, ya que contempla los diversos factores etiológicos y variaciones individuales. Este punto de vista es abordado en 2009 por Tortajada et al. en su estudio basado en datos clínicos de 1397 mujeres que acababan de dar a luz [12].

En él se utilizan variables biológicas y psicosociales para obtener modelos de clasificación basados en Redes Neuronales Artificiales (RNA) para predecir la DPP durante las 32 semanas después del parto. El modelo con mejor precisión que obtuvieron presentaba una tasa de acierto y AUC del 84% y 0.84 respectivamente. Es en este último estudio en el que está basado el presente trabajo.

1.1. Objetivos

Los objetivos principales de este trabajo son los siguientes:

- Revisar el estado del arte en técnicas y métodos para la detección temprana de la DPP.
- Desarrollar nuevos modelos de clasificación para la detección temprana de DPP siguiendo la metodología de minería de datos.
- Evaluar empíricamente los modelos obtenidos.
- Diseñar e implementar una aplicación para dispositivos móviles que integre los modelos para la predicción de la DPP para dos perfiles de usuario: Madres que hayan dado a luz recientemente y personal clínico especializado.

La base de datos que se utilizará en el desarrollo de los modelos de clasificación es la misma que se usó durante el estudio de Tortajada et al. [12]. Puesto que el objetivo final de este trabajo es la creación de una aplicación de ayuda al diagnóstico y detección precoz de la DPP, orientada a personal clínico y a madres, es necesario saber qué diferencias existirán entre los dos modos de funcionamiento o versiones de la misma.

En el caso de la versión para personal clínico, se utilizarán todos los datos clínicos que se disponen, exceptuando una de las variables, para la creación de nuevos modelos de clasificación de la DPP, esperando que presenten resultados parecidos al estudio de Tortajada et al. Los modelos que se refieran a este conjunto de datos, se llamarán durante este trabajo modelos tipo MÉDICO.

En el caso de la versión destinada a madres que han dado a luz recientemente se necesita obtener un modelo sólo con las variables que éstas puedan entender para introducir sus datos como entrada en la aplicación. Puesto que este modelo contemplará menos variables es de esperar una menor precisión de clasificación, pero a cambio podrá advertir a muchas madres sobre la posibilidad de ser población de riesgo. Esto viene ligado con la idea de un sistema de cribado eficaz y coste-eficiente ya que permitiría que, a través de la respuesta a unas simples preguntas, se detecten posibles casos de DPP que de otra manera quedarían sin diagnóstico. Los modelos que se refieran a este conjunto más pequeño de datos durante este trabajo, se les llamará modelos PACIENTE.

Se analizarán los datos y se aplicará una metodología de experimentación para el desarrollo, validación y evaluación de diferentes modelos de clasificación. El siguiente paso consistirá en integrar el clasificador con mejores prestaciones, tanto en su versión para personal clínico, como en la destinada a madres, en un SADC para plataformas móviles Android. Dicho SADC tendrá en cuenta qué tipo de usuario está utilizando la aplicación, y ha de actuar en consecuencia.

En caso de ser personal clínico quien utilice nuestra aplicación, las preguntas podrán ser formuladas con un lenguaje más técnico, y los resultados de clasificación se calcularán en función del mejor modelo de clasificación de tipo MÉDICO.

En caso de ser una madre después del parto quien utilice nuestra aplicación, las preguntas se le han de formular de forma que pueda comprenderlas, con un lenguaje más coloquial, y los resultados de clasificación se calcularán en función del mejor modelo de clasificación de tipo PACIENTE.

En ambas versiones de la aplicación, se deberán presentar los resultados a los usuarios de una forma clara y entendible.

Se pretende pues, poner en manos tanto de personal clínico como de pacientes una herramienta que ayude a prevenir la enfermedad y a detectar la población de riesgo. Por lo tanto, este trabajo abarca todas las etapas de una investigación, además de la transferencia de esta tecnología y conocimiento a una aplicación para dispositivos móviles de fácil utilización y difusión.

2. Materiales y Métodos

A lo largo de esta sección se pretende describir los materiales y metodología empleados para conseguir los objetivos de este trabajo. Estos objetivos, como se ha mencionado anteriormente, son la obtención de modelos de clasificación de la DPP y el posterior desarrollo de una aplicación para dispositivos móviles que puedan usar tanto personal clínico como madres que acaban de dar a luz, mostrando un cuestionario distinto en cada caso.

Se detallará la procedencia y características de los datos utilizados en el entrenamiento de los distintos modelos de clasificación. Para ello, en primer lugar se hará una descripción del estudio prospectivo del que se obtuvo información de unas 1397 mujeres que acababan de dar a luz. En segundo lugar se describirá brevemente las variables independientes que contiene este estudio. Posteriormente se introducirá brevemente las técnicas de Aprendizaje Automático y Reconocimiento de Patrones. Finalmente se enumerarán y describirán los distintos tipos de modelos de clasificación a entrenar con los datos anteriores durante esta investigación, así como los criterios de evaluación de los mismos.

2.1. Descripción del Estudio Prospectivo de donde proceden los datos

Seguidamente se describe el estudio prospectivo del que procede la información que más adelante utilizaremos en la creación de nuestros modelos de clasificación. Los datos provienen del proyecto llamado 'Vulnerabilidad genético-ambiental a la depresión postparto' dirigido por el doctor en psiquiatría Julio Sanjuán Arias y financiado por el Instituto de Salud Carlos III [28]. Toda la información original fue recopilada en siete hospitales generales españoles, en el periodo comprendido entre diciembre de 2003 y octubre de 2004. Todas las participantes eran caucásicas, capaces de leer y responder a los cuestionarios clínicos y ninguna estuvo bajo tratamiento psiquiátrico durante el embarazo. Se excluyó a las mujeres cuyos hijos murieron tras el parto. Este estudio fue aprobado por los Comités de Ética de Investigación Locales, y todas las pacientes dieron su consentimiento informado por escrito. Estos mismos datos se preprocesaron y utilizaron durante la investigación de Tortajada et al. [12], en la que se basa el presente trabajo.

De las 1880 mujeres incluidas inicialmente en el estudio, fueron excluidas 76 debido a que no rellenaron correctamente todas las escalas o cuestionarios. Con las restantes, se realizó un estudio prospectivo en tres instantes de tiempo diferentes: justo después de haber dado a luz, a las 8 y a las 32 semanas después del nacimiento. A las 8 semanas del seguimiento, 1407 (78%) mujeres continuaron en el estudio. A las 32 semanas del seguimiento, 1397 (77.4%) fueron evaluadas. Se comparó las que habían abandonado el seguimiento con las restantes de la muestra que continuaron. Sólo las de clase social más baja aumentaron significativamente en los casos de abandono del seguimiento

(p=0.005). Finalmente, el 11.5% (160) de las mujeres que terminaron el seguimiento a las 8 y 32 semanas tuvieron un episodio de depresión mayor durante los ocho meses siguientes al parto. Por lo tanto, de un número total de 1397 pacientes, se clasificó 160 en la clase positiva y 1237 en la negativa. En la Figura 1 se muestra un diagrama temporal de este estudio prospectivo.

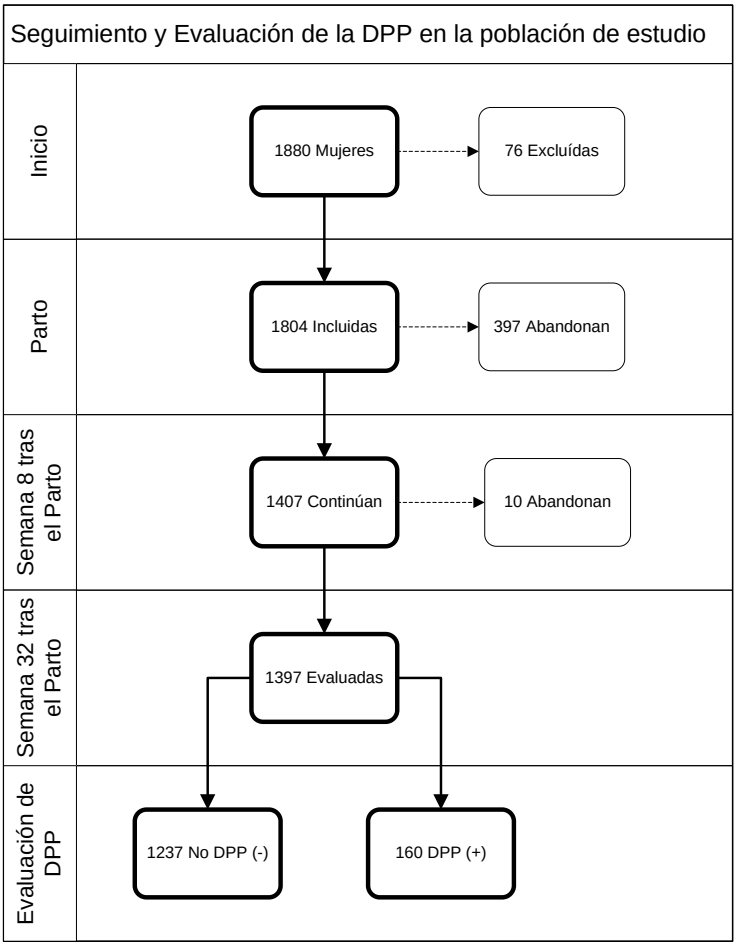


Figura 1: Diagrama temporal del seguimiento y evaluación de la DPP en la población del estudio prospectivo.

El diagrama del protocolo que se utilizó para determinar si las pacientes presentaban DPP se detalla en la Figura 2. Los síntomas depresivos se evaluaron con la puntuación total de la versión española del test EPDS [29] en los tres instantes de tiempo mencionados anteriormente (justo después de haber dado a luz, a las 8 y a las 32 semanas después del nacimiento). Para determinar un episodio de depresión mayor, primero se incluyó sólo aquellas pacientes cuya puntuación en el EPDS fuera mayor o igual a 9 puntos, tras las 8 o 32 semanas después del parto. Estos casos probables, donde la puntuación del EPDS superaba 8 puntos, se evaluaron usando la versión española del DIGS (*Diagnostic Interview for Genetic Studies*) [30] [31] adaptada a la DPP para determinar si la paciente estaba sufriendo un episodio depresivo o no. Las mujeres a

quienes esta última prueba indicaba que sí presentaban una depresión fueron incluidas en la clase positiva, y las que no en la clase negativa. Es decir, en esta última etapa se diagnosticó y separó las mujeres con DPP de las que no la tenían. Todas las entrevistas fueron conducidas por psicólogos clínicos con entrenamiento previo en el DIGS mediante grabaciones de vídeo.

La razón de establecer este protocolo para diagnosticar fehacientemente una DPP es que la prueba EPDS presenta una alta sensibilidad, pero baja especificidad, lo que provoca un elevado número de falsos positivos. De esta manera se consigue detectar a prácticamente todas las enfermas, pero también se identificarían algunas pacientes sanas como casos positivos. Es en este punto donde se utilizó el cuestionario DIGS mediante una entrevista con el psicólogo clínico, siendo esta última prueba el criterio de referencia que determinó finalmente si una mujer estaba sufriendo una DPP o no.

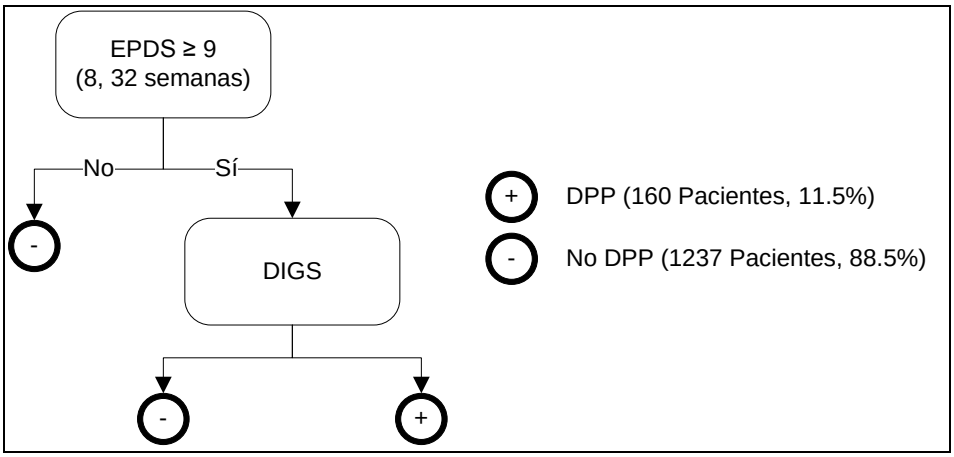


Figura 2: Diagrama de discriminación entre pacientes con y sin DPP en el estudio prospectivo.

Esta discriminación entre casos positivos y negativos de DPP es la que utilizaremos más adelante para entrenar nuestros modelos de clasificación, junto con las variables independientes que describiremos a continuación.

2.2. Descripción de las variables independientes

A continuación se presentará una breve descripción de las variables independientes basadas en el conocimiento actual de la DPP, y utilizadas para desarrollar los modelos de clasificación. Los datos de estas variables se obtuvieron durante el estudio prospectivo descrito anteriormente y contienen información sociodemográfica, psiquiátrica y genética junto con la puntuación del test EPDS justo después del parto.

Todas las participantes rellenaron una entrevista semiestructurada que incluía datos sociodemográficos: edad, nivel de educación, estado civil, número de hijos y situación laboral durante el embarazo. También se tuvo en cuenta el historial personal y familiar de enfermedades psiquiátricas, registrando si existía o no antecedentes psiquiátricos y alteraciones emocionales durante el embarazo como variables binarias.

Respecto a las variables psiquiátricas, el neuroticismo puede definirse como una tendencia perdurable en el tiempo a experimentar estados emocionales negativos. Se midió con la pequeña escala EPQ-N (Eysenck Personality Questionnaire) [32] de 12 ítems, siendo la más comúnmente utilizada en tests de personalidad. En esta investigación se utilizó la versión española validada [33]. Los individuos que obtienen una puntuación alta en esta prueba de neuroticismo son más propensos que la media a experimentar sentimientos de ansiedad, ira, culpa y depresión.

Por otra parte, los eventos traumáticos o eventos vitales estresantes están considerados como factor de riesgo en la literatura y guías clínicas [34]. Se consideró el número de estas experiencias en los tres intervalos de tiempo mencionados en el estudio prospectivo: justo tras el parto, el comprendido desde las 0 hasta las 8 semanas, y el que comprendía las 8 y 32 semanas tras dar a luz. Para ello se utilizó la escala de St. Paul Ramsey [35] [36].

Los síntomas depresivos iniciales fueron evaluados mediante la puntuación del EPDS justo después del parto y también tras las semanas 8 y 32. Las pacientes respondieron al test de 10 ítems a tal efecto, habiendo sido también validado para población española [29]. El mejor umbral de esta escala en la validación española es de una puntuación de 9 para DPP. Sólo el valor inicial de EPDS, es decir, el obtenido justo tras el nacimiento, se utiliza como una variable independiente ya que el objetivo final es prevenir y predecir la DPP dentro de las 32 primeras semanas. Es necesario hacer hincapié en que esta puntuación inicial de EPDS no es la que se utilizó para determinar un episodio de DPP antes de utilizar el cuestionario DIGS en las pacientes del estudio, tal y como se indica en la sección previa.

El apoyo social se midió mediante la versión española de la escala DUKE-UNC [37], que originalmente consiste en un cuestionario de 11 ítems. Este test fue contestado tras el parto, a las 6-8 semanas y en la semana 32. En este trabajo, la variable utilizada es la suma de las puntuaciones obtenidas inmediatamente después del parto más las obtenidas a las 8 semanas. Puesto que se quería predecir el posible riesgo de depresión durante las primeras 32 semanas tras el parto, se descarta la puntuación del test DUKE-UNC de las 32 semanas para la creación de los modelos de clasificación.

Respecto a las variables genéticas, cabe destacar que existen investigaciones donde se relaciona el gen transportador de la serotonina con los cambios de humor después del parto [28]. Es por ello que durante el estudio prospectivo se extrajo ADN genómico de la sangre periférica de las mujeres y se analizaron dos polimorfismos funcionales del gen transportador de la serotonina. La información que se decidió utilizar fue la combinación de genotipos 5-HTT-GC propuesta por Hranilovic [38]. En lo que respecta a esta investigación, lo más relevante en este sentido es que esta información se codificó en una variable categórica con los tres posibles valores siguientes:

- HE: ningún genotipo de baja expresión en cualquiera de los loci
- ME: genotipo de baja expresión en uno de los loci
- LE: genotipo de baja expresión en ambos loci

Finalmente se consideraron otras variables psicosociales y demográficas tales como la edad, el sexo del niño, o el número de miembros de la familia conviviendo con la madre. El nivel más alto de educación alcanzado se codificó en una escala de 3 niveles: alto, medio o bajo. Lo mismo sucede con los ingresos del hogar o nivel económico, donde la escala es: alto, medio, bajo o problemático. La situación laboral durante el embarazo se describe mediante una variable categórica con cuatro posibles valores, siendo éstos los siguientes: empleada, desempleada, ama de casa/estudiante o de permiso.

La Tabla 1 muestra todas las variables clínicas que hemos descrito y las correspondientes abreviaturas que utilizaremos. Hay que tener en cuenta que, como se ha mencionado anteriormente, el objetivo final de este trabajo es, además de la creación de modelos de clasificación, el desarrollo de una aplicación móvil que haga uso de estos clasificadores, con dos tipos de usuarios finales posibles. El primer tipo será personal clínico capacitado para manejar toda esta información, y el otro madres que acaben de dar a luz. Puesto que no es posible que algunos de los datos anteriores sean introducidos por las propias mujeres que acaban de dar a luz, es evidente que los modelos a entrenar en estos casos tendrán que prescindir de esta información. Es por ello que se decidió entrenar dos versiones de cada tipo de modelo de clasificación. La primera se entrenará con todas las variables, a la que llamamos versión ‘MÉDICO’. La segunda la denominaremos versión ‘PACIENTE’, y se excluirán de ella las siguientes variables: cuestionarios EPQ-N de neuroticismo y DUKE-UNC de apoyo social, análisis genético 5-HTT-GC, y número de eventos vitales hasta las 8 y 32 semanas. Los test se excluyen ya que si la madre ha de contestar a demasiados cuestionarios después del parto sin supervisión médica posiblemente no preste atención a las últimas preguntas. Es por eso que se decide dejar únicamente el EPDS en la versión PACIENTE. Es bastante improbable que la madre conozca sus datos genéticos, y por eso también queda fuera la variable 5-HTT-GC. Y finalmente, puesto que se desea que las mujeres utilicen la aplicación durante la primera semana después del parto, los eventos vitales a las 8 y 32 semanas también quedan descartados. Por supuesto, eliminar toda esta información del aprendizaje de los modelos tendrá un coste en sus prestaciones, pero permitirá que se puedan utilizar en el momento adecuado y que las madres puedan responder a preguntas que conozcan y entiendan.

Finalmente, cabe destacar que en todo este trabajo se prescindió de una variable que sí se utilizó durante la investigación de Tortajada [12], la cual es el riesgo médico perinatal

para la madre durante el embarazo. Puesto que existe división de opiniones sobre si esta circunstancia afecta o no en el desarrollo de la DPP, tanto a nivel médico como estadístico, se optó por no incluir la variable que lo representaba en el entrenamiento de los modelos y posterior desarrollo de las aplicaciones.

Variable	Abreviatura
Edad	Edad
Nivel Educativo	N-Educativo
Situación Laboral durante el Embarazo	Sit-Laboral
Nº de personas conviviendo	N-Conviviendo
Alteraciones emocionales durante el embarazo	Alt-Emocion
Antecedentes Psiquiátricos	Antec-Psi
Neuroticismo (EPQ-N)	EPQN
Nº de eventos vitales tras el parto	N-exp-ini
Nº de eventos vitales a las 8 semanas del parto	N-exp-8s
Nº de eventos vitales a las 32 semanas del parto	N-exp-32s
Síntomas depresivos (EPDS Inicial)	EPDS
Nivel Económico	N-Economic
Apoyo social (DUKE, 0 y 8 semanas del parto)	DUKE-suma
5-HTT-GC (Análisis genético)	5-HTT-GC
Sexo del bebé	Sexo

Tabla 1: Variables independientes y sus abreviaturas utilizadas durante el estudio

A partir de los datos de estas variables independientes se plantea un problema de clasificación entre las madres que finalmente desarrollaron DPP y las que no. Este tipo de problemas son abordados por las ramas de la inteligencia artificial referentes al Aprendizaje Automático y el Reconocimiento de Patrones. En los siguientes apartados se introducirán estas aproximaciones, además de presentar los distintos tipos de modelos de clasificación que se utilizarán durante este trabajo.

2.3. Descripción de Aprendizaje Automático y Reconocimiento de Patrones

Un diagnóstico médico es un proceso cognitivo en el que el médico intenta identificar una enfermedad en un paciente. El diagnóstico está basado en una serie de datos que sirven como información de entrada para llegar a un diagnóstico. Por lo tanto, este proceso puede ser considerado como un problema de clasificación. El campo del Reconocimiento de Patrones (RP) se refiere al descubrimiento de características en los datos para clasificarlos en diferentes categorías [39]. Es decir, el RP se refiere a aquellos modelos matemáticos o que mediante estructuras de datos más o menos complejas, consiguen etiquetar o reconocer una nueva muestra de datos en la clase a la que pertenece.

Los modelos de RP se desarrollan normalmente mediante técnicas de Aprendizaje Automático (AA), las cuales proveen mecanismos matemáticos y computacionales para inferir conocimiento de los datos específicos de un dominio concreto.

El ciclo de vida de un problema de RP basado en AA puede dividirse en dos fases principales: entrenamiento y reconocimiento. Durante la fase de entrenamiento se utiliza un conjunto de datos para construir el modelo de RP. Es en esta fase donde se ajusta un modelo adaptativo para obtener la mejor generalización posible, y de ese modo resolver nuevos casos durante la fase de reconocimiento. Una vez el modelo está listo, se puede incorporar en un SADM para ayudar en futuras observaciones a reconocer dichos patrones.

El problema general del AA se suele describir ayudándose de un proceso generador de observaciones aleatorias llamado s, el cual es obtenido siguiendo un proceso de dos etapas. Primero, un generador produce vectores de características aleatorias $x \in X^D$ siguiendo una función de distribución de probabilidad $p(c|x)$, produciendo muestras tales como $s_i = \{x_i, c_i\}$ con probabilidad $p(x_i, c_i)$, donde $p(x_i, c_i) = p(x_i)p(c_i|x_i) = p(c_i)p(x_i|c_i)$. El valor $p(c)$ es conocido como la probabilidad a priori de la clase y al valor $p(x|c)$ se le llama la probabilidad condicional de la clase, o simplemente la probabilidad condicional.

El objetivo del AA es desarrollar una regla de decisión o modelo M que asigne un vector de características a una clase. Por lo tanto, un modelo es una asignación $M: X^D \rightarrow C$. Generalmente, el modelo se puede definir como una función parametrizada $\hat{c} = f_M(x, \alpha), \alpha \epsilon \Lambda$ que intenta aproximar el valor de c. Es por lo tanto posible medir las consecuencias de aproximar $\hat{c}$ dado x mediante una función de pérdida (*loss function*)

$$L(c, \hat{c}) = \begin{cases} 0 \; si \; \hat{c} = c \\ 1 \; si \; \hat{c} \neq c \end{cases}$$

Esta función es también conocida como la función de pérdida 0-1. Gracias a ella es posible calcular el riesgo condicional del modelo M, expresado como

$$R(\hat{c}|x) = \sum_{c \in C} L\big(c, f_M(x;\, \alpha)\big) p(c|x)$$

Se define como regla de decisión óptima aquella que consigue una mínima probabilidad de error. Si las probabilidades a priori y las probabilidades condicionales de la clase son conocidas, entonces la regla de decisión óptima es la regla de decisión Bayesiana,

$$\hat{c}^* \leftarrow arg \min_{c \in C} R(\hat{c}|x)$$

Cuando se asume la función de pérdida 0-1, entonces el riesgo condicional es el error de probabilidad medio, el cual se puede expresar como

$$R(\hat{c}|x) = 1 - p(\hat{c}|x)$$

Sin embargo, la situación más común es que estas distribuciones de probabilidad sean desconocidas. Aún así, se pueden aproximar mediante el conjunto de observaciones $S = \{(x_1, c_1), (x_2, c_2), \dots, (x_N, c_N)\} \in (X^D \times C)$ que supuestamente concuerda con $p(x, c)$. El principal objetivo del AA es encontrar, basándose en S, una función o modelo M cuyo riesgo sea lo más cercano posible a $R(\hat{c}^*)$ ayudándose de la regla delta de Kronecker.

Resumiendo, durante el proceso de entrenamiento de RP, se utilizan algoritmos de AA que ajustan los parámetros del modelo en función de la minimización del riesgo condicional del modelo. Estos algoritmos de entrenamiento dependerán del clasificador concreto que se esté utilizando, pero sus fundamentos teóricos son los mismos en todos.

Es importante mencionar que existen los parámetros del modelo que se ajustan durante el proceso de entrenamiento y los hiperparámetros. Podemos definir los hiperparámetros como aquellas variables del modelo que se han de ajustar 'manualmente'. Este ajuste manual o *tuning* se suele hacer midiendo el rendimiento del modelo en un conjunto de datos independiente al de entrenamiento, conocido normalmente por conjunto de validación. Se aplican las mismas reglas de minimización del error del riesgo condicional para entrenar el modelo dentro de un rango permitido de hiperparámetros y se escoge aquel modelo que mejor rendimiento presenta en el conjunto de validación. A este proceso de búsqueda de la mejor combinación de hiperparámetros se le suele llamar red de búsqueda o *grid search*.

2.4. Modelos de clasificación a utilizar durante la experimentación

Una vez tenemos claras y bien definidas las variables independientes que disponemos, y los fundamentos del AA y el RP, seguidamente se pasará a enumerar y explicar brevemente los fundamentos teóricos de los distintos tipos de modelos de clasificación que se utilizaron durante la experimentación de este trabajo. Dicha experimentación se compone de un proceso de aprendizaje, validación y evaluación de los clasificadores que se detallará más adelante. No obstante, es necesario comprender los principios de estos modelos y el comportamiento que hemos de esperar de ellos.

Tal y como dijo el famoso estadista George E. P. Box: "Todos los modelos son erróneos, pero algunos son útiles" [40]. Un modelo no es más que el intento matemático de representar la realidad, pero ésta es a menudo demasiado compleja para ser representada sin errores o sesgos. El ámbito clínico es uno de los campos donde más se cumplen estas afirmaciones, ya que los problemas que nos encontramos en él contienen variables difíciles de modelar y otras que ni tan siquiera se han descubierto. El caso de la DPP no es menos, y es evidente que las variables presentadas en el apartado anterior no son suficientes para obtener un clasificador perfecto, pero sí que son útiles para obtener unos resultados estadísticamente aceptables. Además, puesto que ante un problema de esta índole no se conoce a priori qué tipo de clasificador modelará mejor nuestra realidad, se decidió experimentar con distintos tipos de ellos.

Así pues, durante la realización de este trabajo se experimentó con 6 tipos diferentes de clasificadores: Naïve Bayes, Regresión Logística, SVM (Support Vector Machines), RNA (Redes Neuronales Artificiales), K-NN (K-Nearest Neighbours) y Árboles de Decisión. Al ser Naïve Bayes el más sencillo de todos estos métodos, se utilizó como referencia de rendimiento mínimo en la evaluación, tomando como criterio de este estudio presentar sólo los resultados de aquellos métodos más sofisticados que mejorasen esta referencia. No sucedió así con los Árboles de Decisión [41] ni con K-NN [42], por lo que ni los resultados ni fundamentos teóricos aparecen aquí.

Toda la parte experimental se llevó a cabo en un PC con procesador AMD Athlon™ 64 Processor 3200+ con una velocidad de reloj de 2 GHz, 2.50 GB de memoria RAM, bajo el sistema operativo Windows 7 Professional de 32 bits. El software utilizado fue la versión 2013a de Matlab (Mathworks).

2.4.1. Naïve Bayes

El clasificador *Naïve Bayes* [43], se basa en el teorema de Bayes asumiendo independencia entre las variables independientes o predictores. Es un modelo fácil de construir y sin ningún hiperparámetro a estimar. A pesar de su simplicidad, en muchas ocasiones muestra un rendimiento sorprendentemente bueno y es ampliamente usado ya que en algunos problemas mejora los resultados de clasificación obtenidos con métodos más sofisticados.

El teorema de Bayes provee un método para calcular la probabilidad a posteriori de la clase a la que pertenece el objeto a clasificar. El clasificador Naïve Bayes asume que el efecto del valor de un predictor (x) en una clase (c) es independiente de los valores de otro predictor. Esta asunción se llama independencia condicional de la clase. En la Figura 3 aparece la ecuación básica de la probabilidad de clasificar una nueva muestra como clase c, dados unos datos de entrenamiento o predictores X, con Naïve Bayes. A la hora de clasificar una nueva muestra, se escoge de entre todas las clases aquella que mayor probabilidad tenga según dicha ecuación. Como podemos ver, el resultado dependerá única y exclusivamente de los datos de entrenamiento, no existiendo ningún hiperparámetro a estimar.

$$P(c|X) = \prod_{i=1}^{n} P(x_i|c) \cdot P(c)$$

Figura 3: Ecuación básica de la probabilidad de clasificar una muestra como clase c, dados unos predictores X, con Naïve Bayes.

2.4.2. Regresión Logística

Los modelos de *Regresión Logística* [44], se utilizan cuando la variable dependiente es binaria, es decir, predice la probabilidad de un resultado que sólo puede tener dos valores posibles. Este es el caso que nos ocupa, ya que se busca distinguir entre madres con y sin DPP. Las variables independientes (X), también llamadas predictores, pueden aparecer de forma numérica o categórica.

Una regresión lineal no es adecuada para predecir el valor de una variable binaria por las dos razones siguientes. La primera es que una regresión lineal predecirá valores fuera del rango aceptable (por ejemplo, la predicción de probabilidades fuera del rango [0, 1]). La segunda es que, puesto que la variable de respuesta (Y) sólo puede tener uno de los dos valores posibles para cada muestra, los residuos no se distribuye normalmente alrededor de la línea prevista.

Por otro lado, una regresión logística produce una curva sigmoide, que se limita a valores entre 0 y 1. Es similar a una regresión lineal, pero la curva se construye utilizando el logaritmo neperiano de los *'odds ratio'* de la variable de destino, en lugar de la probabilidad. Por otra parte, las variables independientes no necesitan seguir una distribución normal o tener igualdad de varianza en cada grupo.

En la Figura 4 se puede apreciar las diferencias descritas entre un modelo de regresión lineal *y*, frente al de regresión logística *p*.

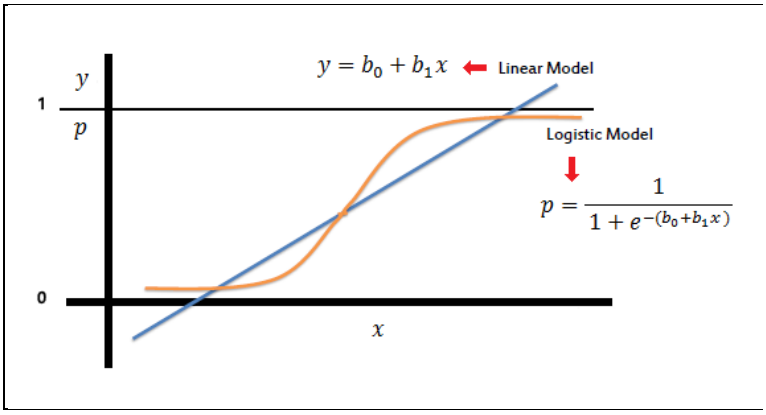


Figura 4: Diferencias entre un modelo de Regresión Lineal y otro de Regresión Logística [45].

Si queremos que el modelo proporcione la probabilidad p_i de pertenecer a cada una de las clases, debemos transformar la variable respuesta de algún modo para garantizar que la salida prevista esté entre cero y uno. Para ello tomamos la función de distribución logística de la ecuación de la Figura 5. La función logística anterior se puede convertir en un modelo lineal usando la transformada *logit* g_i de la ecuación de la Figura 6 [46]. La probabilidad logarítmica es usada para estimar los coeficientes de regresión (β_i) del modelo. Así pues, los valores exponenciales de los coeficientes de la regresión dan como resultado el valor de *odds ratio*, el cual refleja el efecto de la variable de entrada como un factor de riesgo o protector.

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_i X_i)}}$$

Figura 5: Ecuación básica de la salida de una Regresión Logística

$$g_i = \ln \frac{p_i}{1 - p_i} = \beta_0 + \beta_i X_i$$

Figura 6: Ecuación de la transformación *logit* de la salida de una Regresión Logística.

Para evitar el sobreajuste y asegurar la relevancia de una variable independiente en este trabajo, se comprueban que todos los valores de los coeficientes β_i sean mayores a 0.01. De no ser así, se realiza una nueva regresión logística sin esas variables y se compara los resultados con la anterior, decidiendo finalmente la conveniencia de incluir dichas variables o no en el modelo.

Un hecho a destacar es que los modelos de regresión logística más comunes están limitados a relaciones lineales entre variables dependientes e independientes. Esta limitación se puede superar usando funciones de expansión base o mediante el método de los polinomios fraccionales [23]. Los modelos de redes neuronales también pueden superar esta limitación. Por lo tanto, las relaciones lineales entre variables independientes y dependientes pueden ser encontradas en ambos modelos. Mientras que las interacciones no lineales sólo podrán aparecerán en el modelo conexionista, es decir, las redes neuronales, tal y como se explicará más adelante. El uso de funciones de expansión base y el método de los polinomios fraccionales quedan fuera del alcance de este trabajo.

2.4.3. *Support Vector Machines* (SVM)

En *Support Vector Machines* (SVM) [47], el proceso de clasificación se realiza mediante el hiperplano que maximiza el margen entre dos clases en los datos de entrenamiento. El margen se define como la distancia perpendicular mínima entre dos puntos de cada clase al hiperplano separador. Dicho hiperplano se ajusta durante el proceso de aprendizaje con los datos de entrenamiento o predictores X. De entre estos predictores, se seleccionan los vectores que definen el hiperplano, los cuales son llamados '*support vectors*'. En la Figura 7 aparece un ejemplo con la idea básica que persigue SVM en dos dimensiones. Extrapolándolo a nuestro caso, y suponiendo que sólo tuviéramos dos variables independientes x_1 y x_2, los puntos verdes podrían representar las mujeres sin DPP, mientras que los puntos rojos las que finalmente sí desarrollaron la enfermedad. En un caso ideal un plano podría separar estas dos clases ayudándose de un margen de distancia a dicho plano. Si pensamos en las 15 variables independientes que disponemos, tendremos que tener en cuenta que nuestro espacio tiene 15 dimensiones (más las añadidas por la codificación de las variables categóricas). El hiperplano será de dimensión D-1, donde D es el número de dimensiones o variables independientes.

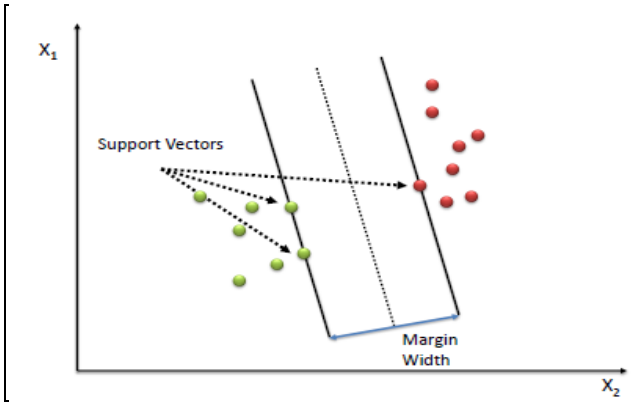


Figura 7: Idea básica de hiperplano en SVM con dos dimensiones [48].

Tal y como se aprecia en la Figura 8, los hiperplanos correspondientes a $\vec{w} \cdot \vec{x} + b = -1$ y $\vec{w} \cdot \vec{x} + b = 1$ son los hiperplanos frontera y definen el margen. La distancia entre los dos hiperplanos frontera es el margen, el cual es igual a $2/\|w\|$. Se puede demostrar que, dados unos datos de entrenamiento, maximizar el margen de separación entre dos clases tiene el efecto de reducir la complejidad del clasificador y por lo tanto optimizar la generalización. El hiperplano óptimo corresponde a aquel que minimiza el error de entrenamiento y, al mismo tiempo, tiene el máximo margen de separación entre las dos clases. Para generalizar los casos donde los límites de decisión no son linealmente separables, SVM proyecta los datos de entrenamiento en otro espacio de dimensionalidad más alta. Si la dimensionalidad del nuevo espacio es suficientemente alta, los datos siempre serán linealmente separables. Para evitar tener que realizar una proyección explícita en un espacio dimensional mayor se utiliza una función *kernel* (K). Esta función K es la que transforma implícitamente los datos a este espacio dimensional mayor para hacer posible la separación lineal de las clases. K puede ser de tipo polinomial, de base radial Gaussiana, o perceptrón sigmoide [49].

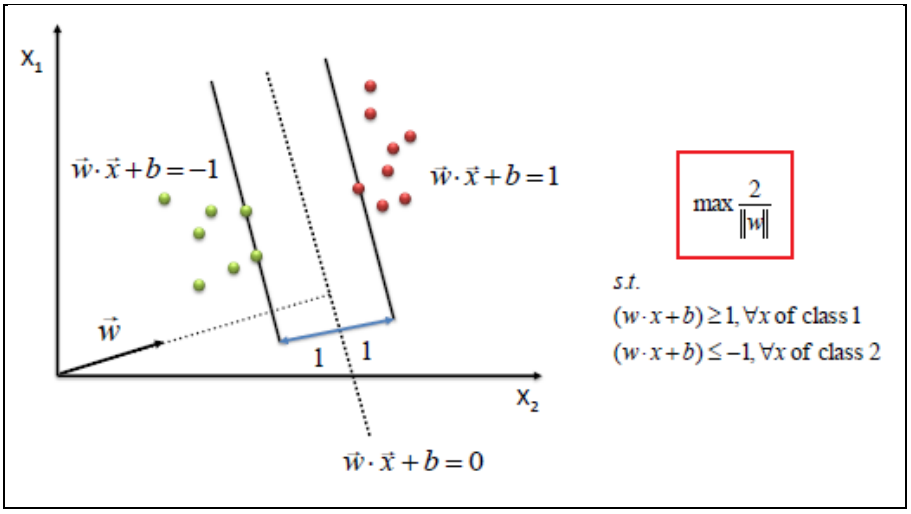


Figura 8: Ejemplo de hiperplano en SVM y sus parámetros básicos con dos dimensiones [48].

2.4.4. Redes Neuronales Artificiales (RNA)

Las *Redes Neuronales Artificiales* (RNA) se inspiran en sistemas biológicos en los que un gran número de unidades simples trabajan en paralelo para realizar tareas más complejas. Estas redes están hechas de muchos procesadores simples (neuronas o unidades) basadas en el perceptrón de Rosenblatt [50]. Un perceptrón da una combinación lineal, y, de los valores de sus d entradas x_i, más un valor de *bias* o sesgo b. Los valores de los pesos de cada entrada se expresa mediante w_i.

La salida se calcula aplicando una función de activación a la combinación lineal de los pesos y las entradas. Normalmente, la función de activación es una identidad, una logística o una tangente hiperbólica. Puesto que estas funciones son monotónicas, la forma $f(\sum_{i=1}^{d} x_i w_i + b)$ es una función lineal discriminante [51]. La Figura 9 muestra la ecuación y el esquema básico de un perceptrón de estas características.

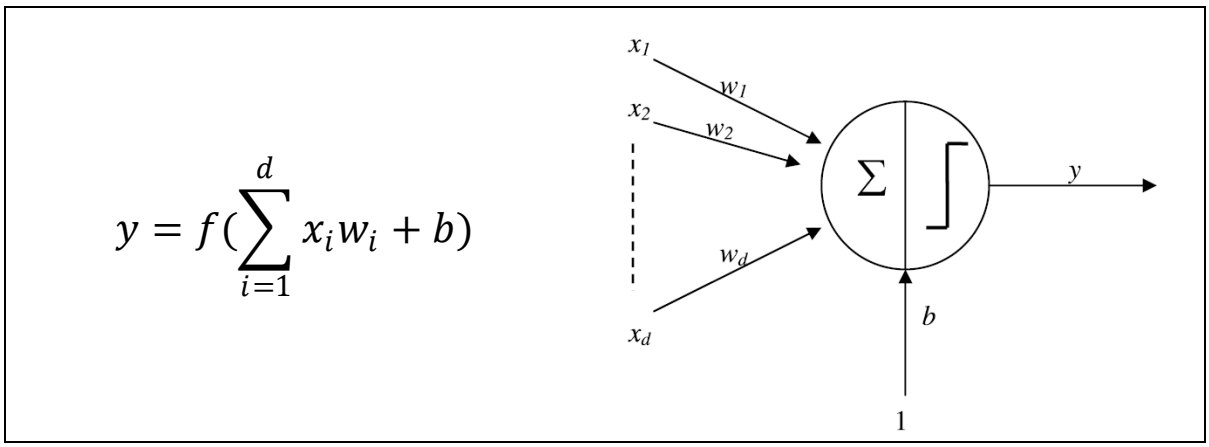


Figura 9: Ecuación y esquema básico de un Perceptrón con 'd' entradas [49].

Una única unidad tiene una habilidad computacional limitada, pero un grupo de neuronas interconectadas tienen una capacidad de adaptación muy potente, y la habilidad de aprender funciones no lineales, pudiendo modelar relaciones complejas entre entradas y salidas. Por lo tanto, se pueden construir funciones más generales al crear redes con sucesivas capas de unidades de procesamiento, con conexiones desde cada unidad de una capa a todas las unidades de la capa siguiente. Un perceptrón multicapa *feed-forward* consiste en una capa de entrada con una unidad para cada variable independiente, una o dos capas ocultas más de perceptrones, y la capa de salida para la variable dependiente (en el caso de un problema de regresión), o las posibles clases (en el caso de un problema de clasificación). Se le llama perceptrón multicapa *feed-forward* totalmente conectado cuando toda unidad de cada capa recibe como entrada la salida de cada unidad de su capa precedente y la salida de cada unidad es enviada a cada unidad de su siguiente capa. La Figura 10 muestra un ejemplo de RNA *feed-forward* con dos capas ocultas y una única unidad de salida.

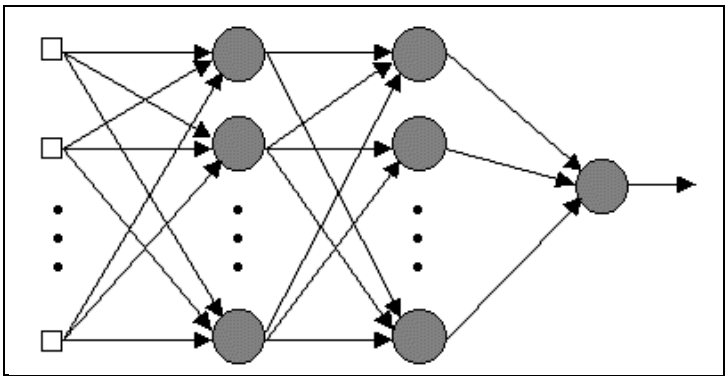


Figura 10: Ejemplo de RNA con dos capas ocultas [49].

Se utilizó como función de activación de la salida de todas las unidades la tangente hiperbólica la cual está dentro del rango [-1, 1]. Puesto que en este trabajo se considera la PPD como una variable binaria dependiente, la función de activación de la unidad de salida se escaló posteriormente en el rango [0 1].

Aunque estos modelos, y las RNAs en general, muestran una gran potencia predictiva comparada con otras aproximaciones tradicionales, se les ha etiquetado como métodos de ‘caja negra’. Esto es así ya que no poseen un gran poder descriptivo sobre la influencia relativa de las variables independientes en el proceso predictivo. Esta falta de poder explicativo es un problema para conseguir una interpretación de la influencia de cada variable independiente en la PPD.

Otra de las características que presentan este tipo de modelo es que normalmente se necesita hacer un barrido de combinaciones sobre el número de capas, número de unidades por capa e inicialización aleatoria de los pesos para encontrar un modelo que se ajuste bien al problema en cuestión. En nuestro caso, y tal como se explicará en el apartado de experimentación, se utilizaron perceptrones multicapa *feed-forward* totalmente conectados con una y dos capas ocultas. Una vez presentados todos los tipos de clasificadores a utilizar durante este trabajo, el siguiente apartado explicará los criterios de evaluación de los modelos de clasificación con los que se experimentó.

2.5. Criterios de evaluación de los modelos de clasificación

Llegados a este punto queda claro que tenemos a nuestra disposición una base de datos con determinadas variables independientes relativa a mujeres que acababan de dar a luz, y donde cada paciente había sido clasificada por un criterio de referencia en la clase positiva (DPP) o negativa (NO DPP). La prevalencia de DPP en esta base de datos de 1397 pacientes es de un 11.5%. A partir de todos estos datos se desea entrenar los modelos de clasificación descritos en el apartado anterior.

La evaluación de los modelos se realizó utilizando validación *hold-out*, donde las muestras o pacientes fueron escogidas aleatoriamente para formar los conjuntos de entrenamiento, validación y evaluación. Para obtener un buen error de estimación de los modelos predictivos, se dividió la base de datos en tres conjuntos o *datasets* diferentes: El conjunto de entrenamiento con 1006 pacientes (72%), el conjunto de validación con 112 pacientes (8%), y el conjunto de test con 279 pacientes (20%). La proporción de estas divisiones se pueden apreciar en la Figura 11.

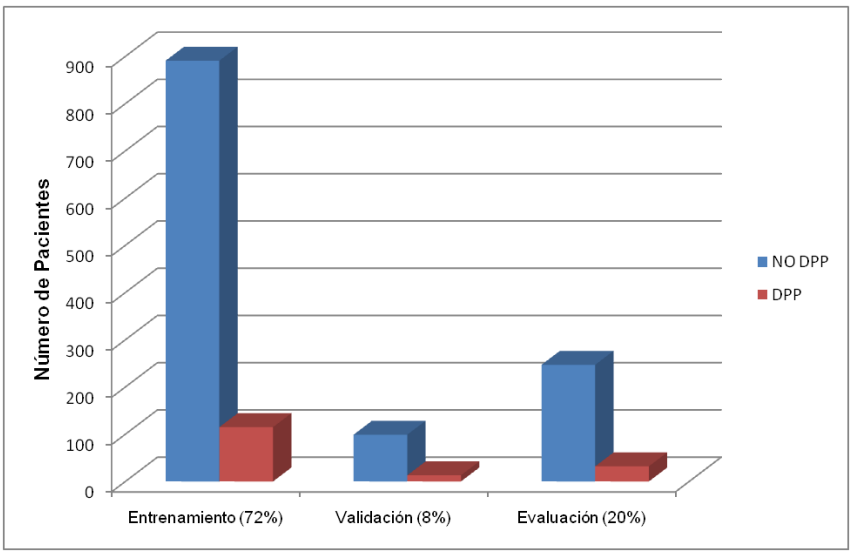


Figura 11: Gráfico de barras con el número de muestras y su clase en los conjuntos de entrenamiento, validación y evaluación.

Cada partición siguió la prevalencia de la base de datos original, tal como se muestra en la Tabla 2. Los hiperparámetros de los distintos modelos se seleccionaron empíricamente usando el conjunto de validación, para más adelante ser evaluados con el conjunto de evaluación. De este modo se evitó el sobreajuste usando el conjunto de validación durante el proceso de aprendizaje.

DATASET	NO DPP		DPP		TOTAL
Entrenamiento	(88.6%)	891	(11.4%)	115	1006
Validación	(88.4%)	99	(11.6%)	13	112
Evaluación	(88.5%)	247	(11.5%)	32	279
TOTAL		1237		160	1397

Tabla 2: Número de muestras por clase en cada partición de la base de datos original.

En aquellos modelos donde no existe ajuste de hiperparámetros, como por ejemplo la Regresión Logística o Naïve Bayes, los conjuntos de entrenamiento y validación se unen para el proceso de entrenamiento.

Es evidente que una buena prueba diagnóstica es la que ofrece resultados positivos en enfermos, y negativos en sanos. En nuestro caso, se nos plantea conseguir una prueba dicotómica, es decir, una prueba que clasifica a cada paciente como sana (NO DPP) o enferma (DPP) en función de que el resultado de la prueba sea positivo o negativo. Asignaremos el resultado positivo a la presencia o riesgo de DPP, y el negativo a su ausencia.

En un estudio de estas características, los datos obtenidos permiten clasificar a las pacientes en cuatro grupos, según una matriz de confusión como la que se muestra en la Tabla 3. En las filas encontramos el resultado de la prueba diagnóstica, en nuestro caso el resultado del modelo de clasificación. En las columnas aparece el estado real de las pacientes, que en esta investigación se obtiene de la prueba de referencia explicada anteriormente durante el estudio prospectivo a partir del DIGS. Al enfrentar filas y columnas, el resultado de la prueba puede ser correcto (verdaderos positivos y negativos) o incorrecto (falsos positivos y negativos).

	Verdadero Diagnóstico	
Resultado del clasificador	**Enferma (DPP)**	**Sana (NO DPP)**
Positivo	Verdaderos Positivos (VP)	Falsos Positivos (FP)
Negativo	Falsos Negativos (FN)	Verdaderos Negativos (VN)

Tabla 3: Matriz de confusión de un clasificador. Relación entre el resultado de una prueba diagnóstica y la presencia o ausencia de DPP.

A partir de la matriz de confusión anterior, para cada modelo de clasificación desarrollado indicaremos las medidas que a continuación se detallan. De este modo nos podremos hacer una idea de la validez y seguridad que nos ofrecen. Todas las fórmulas de estas medidas aparecen en la Tabla 4 y están documentadas en muchos trabajos y estudios, entre ellos el de Pita Fernández et al. [52].

Sensibilidad (SEN): Se define como la probabilidad de clasificar correctamente a un individuo enfermo, es decir, la probabilidad de que para un sujeto enfermo el clasificador dé un resultado positivo. Es, por lo tanto, la capacidad del test para detectar la enfermedad y está relacionada con la validez de la prueba diagnóstica.

Especificidad (ESP): Se define como la probabilidad de clasificar correctamente a un individuo sano, es decir, la probabilidad de que para un sujeto sano se obtenga un resultado negativo. Es, por lo tanto, la capacidad del test para detectar a los sujetos sanos y también está relacionada con la validez de la prueba diagnóstica.

Valor Predictivo Positivo (VPP): Es la probabilidad de padecer la enfermedad si se obtiene un resultado positivo en el test. Esta medida está relacionada con la seguridad de la prueba diagnóstica. No obstante, en nuestro trabajo esta medida sólo aparecerá a título informativo, ya que las decisiones sobre qué clasificadores se utilizarán finalmente se tomarán en base a la Sensibilidad y Especificidad.

Valor Predictivo Negativo (VPN): Es la probabilidad de que un sujeto esté realmente sano ante un resultado negativo en el test.

Razón de Verosimilitud Positiva (RV+): Se calcula dividiendo la probabilidad de un resultado positivo en los pacientes enfermos entre la probabilidad de un resultado positivo entre los sanos. Esta medida está relacionada con las razones de probabilidad, no dependiendo de la prevalencia de la enfermedad de estudio. Mide cuánto más probable es un resultado concreto (positivo o negativo) según la presencia o ausencia de enfermedad.

Razón de Verosimilitud Negativa (RV-): Se calcula dividiendo la probabilidad de un resultado negativo en presencia de enfermedad entre la probabilidad de un resultado negativo en ausencia de la misma.

$SEN = \frac{VP}{VP + FN}$	$VPP = \frac{VP}{VP + FP}$	$RV+= \frac{SEN}{1 - ESP}$
$ESP = \frac{VN}{VN + FP}$	$VPN = \frac{VN}{VN + FN}$	$RV-= \frac{1 - SEN}{ESP}$

Tabla 4: Medidas de precisión de un clasificador

Durante la presentación de los resultados de cada modelo, también se indicará la *Accuracy* o tasa de aciertos sobre el total de las muestras. Además se incluirá el intervalo de confianza con un nivel de significancia $\alpha = 5\%$.

Existe una dificultad intrínseca en la naturaleza de nuestro problema puesto que la base de datos está desbalanceada [53]. Esto significa que los casos positivos (11%) están infrarrepresentados en comparación con los negativos (89%). Por lo tanto, con esta prevalencia en las muestras negativas, un simple clasificador consistente en asignar la clase más probable *a priori* a una nueva muestra alcanzaría una precisión de alrededor del 89%. Sin embargo, su sensibilidad sería nula.

El principal objetivo es obtener modelos predictivos con una buena sensibilidad y especificidad. Ambas medidas dependen de la precisión en las muestras positivas o enfermas (VP), y la precisión en las negativas (VN). Hay que tener claro que incrementar los VP será a costa de decrementar los VN.

La relación entre estas medidas puede ser mostrada mediante una curva ROC [54]. Una curva ROC (acrónimo de *Receiver Operating Characteristic*, o Característica Operativa del Receptor) es una representación gráfica de la sensibilidad frente a (1 – especificidad) para un sistema clasificador binario según se varía el umbral de discriminación. El área bajo la curva (AUC) es una medida independiente de la prevalencia de la enfermedad en la población de estudio, y refleja la bondad del test para discriminar pacientes con y sin la enfermedad a lo largo de todo el rango de puntos de corte o umbrales posibles [55]. Este área posee un rango de valores comprendido entre 0.5 y 1, donde 1 representa un valor diagnóstico perfecto y 0.5 es una prueba sin capacidad discriminatoria diagnóstica. Es

decir, si AUC para una prueba diagnóstica es 0.8 significa que existe un 80% de probabilidad de que el diagnóstico realizado a una persona enferma sea más correcto que el de una persona sana escogida al azar. La Figura 12 muestra algunos ejemplos de curvas ROC junto su AUC y el valor diagnóstico que podríamos esperar de cada una de ellas. Por esto, el AUC nos sirve también para evaluar la calidad de nuestros clasificadores, y cuanto mayor sea su valor, mayor capacidad discriminatoria podremos decir que tiene.

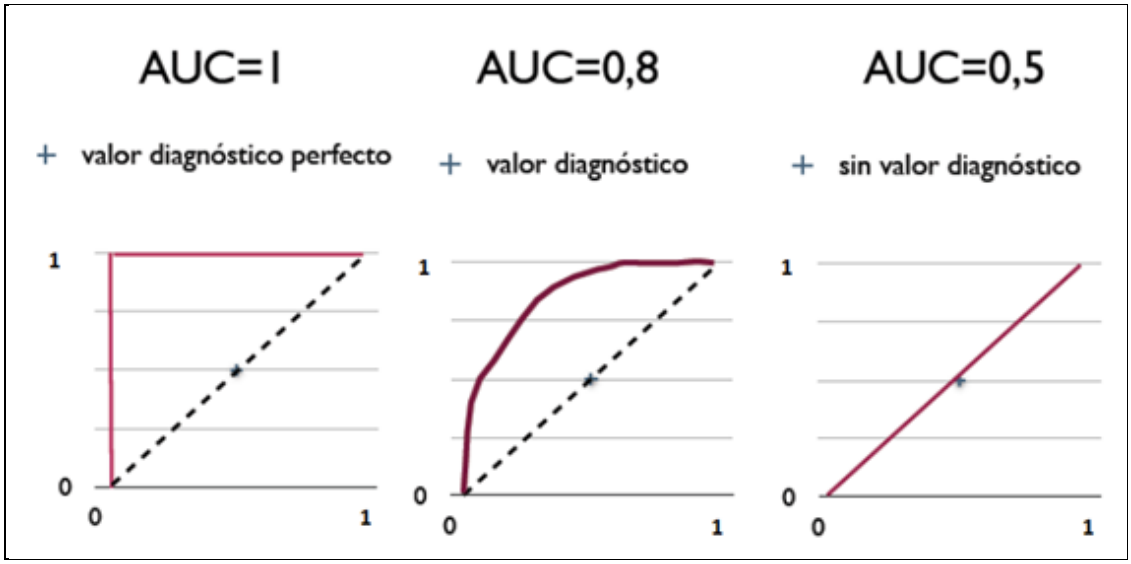


Figura 12: Ejemplos de curvas ROC junto su AUC y valor diagnóstico [56].

Por otro lado, la relación entre sensibilidad y especificidad también puede ser analizada mediante una media geométrica a la que llamaremos G, siendo su fórmula la siguiente:

$$G = \sqrt{SEN \cdot ESP}$$

G es una función comprendida entre 0 y 1, y alcanza valores altos sólo si sus dos operandos, sensibilidad y especificidad, son elevados y están en equilibrio, tal y como se puede ver en la Figura 13. De este modo, utilizando la G para evaluar el modelo simple que habíamos mencionado anteriormente, que siempre asigna la clase con la máxima probabilidad a priori, obtendríamos un valor de G de 0. Esto significa que el modelo es el peor que podemos obtener. Por el contrario, un clasificador ideal, el cual identifica a todos los sanos y enfermos como tales, obtendrá una sensibilidad y especificidad de 1, y por ende, una G también con su valor máximo de 1.

Será esta última medida descrita (G), la que se utilizará para seleccionar los mejores clasificadores, ya que nos relaciona de una manera eficaz y comprensible la sensibilidad y especificidad de nuestra prueba diagnóstica, en nuestro caso el resultado del modelo de clasificación.

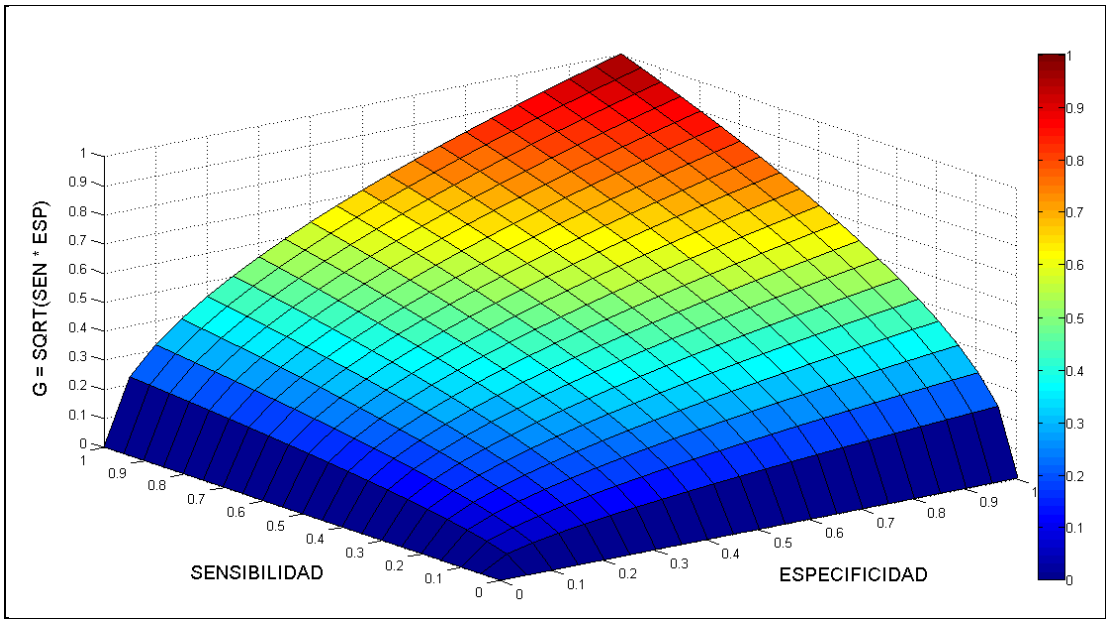


Figura 13: Superficie 3D de la función G: Media geométrica entre Sensibilidad y Especificidad que utilizaremos como criterio de evaluación de los modelos de clasificación.

Si comparamos la función G con la tasa de aciertos o *Accuracy* teniendo en cuenta que en nuestra base de datos las clases están descompensadas, es fácil llegar a la conclusión de que la *Accuracy* no es una buena medida para la evaluación final de nuestros modelos. En la Figura 14 está representada en una superficie 3D la precisión o tasa de aciertos frente al número de aciertos en las distintas clases de nuestro conjunto de datos. Se puede ver como los ejes X e Y de la figura presentan una escala distinta, ya que en la clase DPP sólo existen 160 mujeres, mientras que en la NO-DPP tenemos 1237.

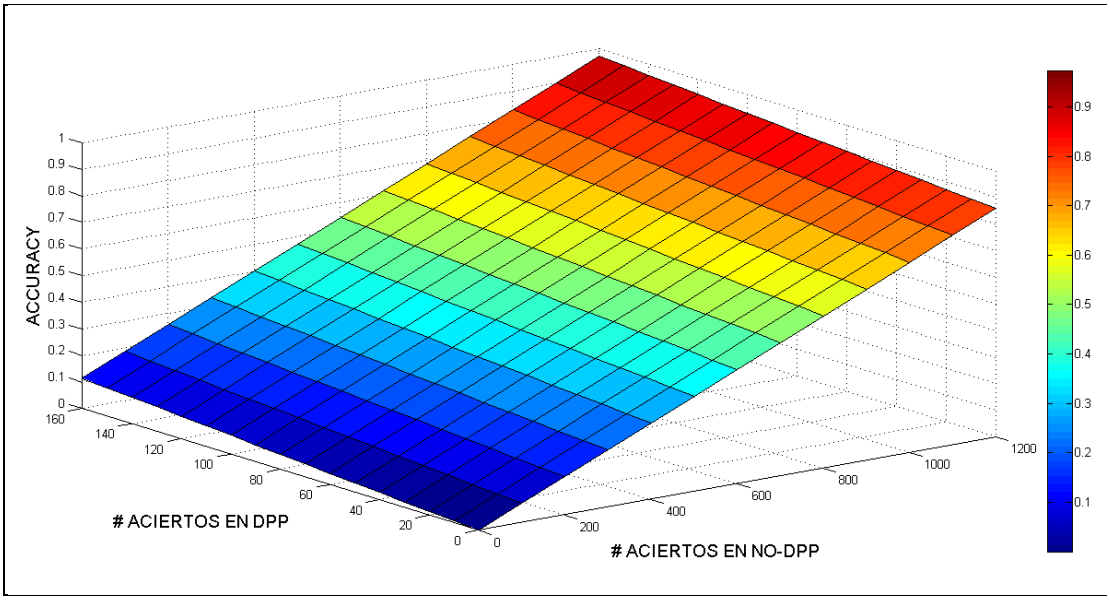


Figura 14: Superficie 3D de la tasa de acierto (*Accuracy*) según el número de aciertos en las distintas clases de nuestro conjunto de datos.

Este desbalanceo provoca que, ante un 100% de aciertos en las mujeres sanas, y un 0% en las enfermas, tengamos una tasa de acierto del 89%. Si fuera a la inversa, esta tasa de aciertos sería del 11%. Esto es fácil verlo si nos fijamos en las esquinas de la función representada en la figura anterior, la cual es un simple plano en el espacio 3D.

Por otro lado, y ante los mismos ejes X e Y que representan el número de aciertos en las distintas clases de nuestra base de datos, en la Figura 15 se aprecia como la función G se comporta correctamente. Se aprecia como balancea bien la relación entre sensibilidad y especificidad que deseamos encontrar. Es por eso que consideramos que G es una medida muy robusta midiendo el rendimiento de clasificadores, incluso si el número de clases está desbalanceado.

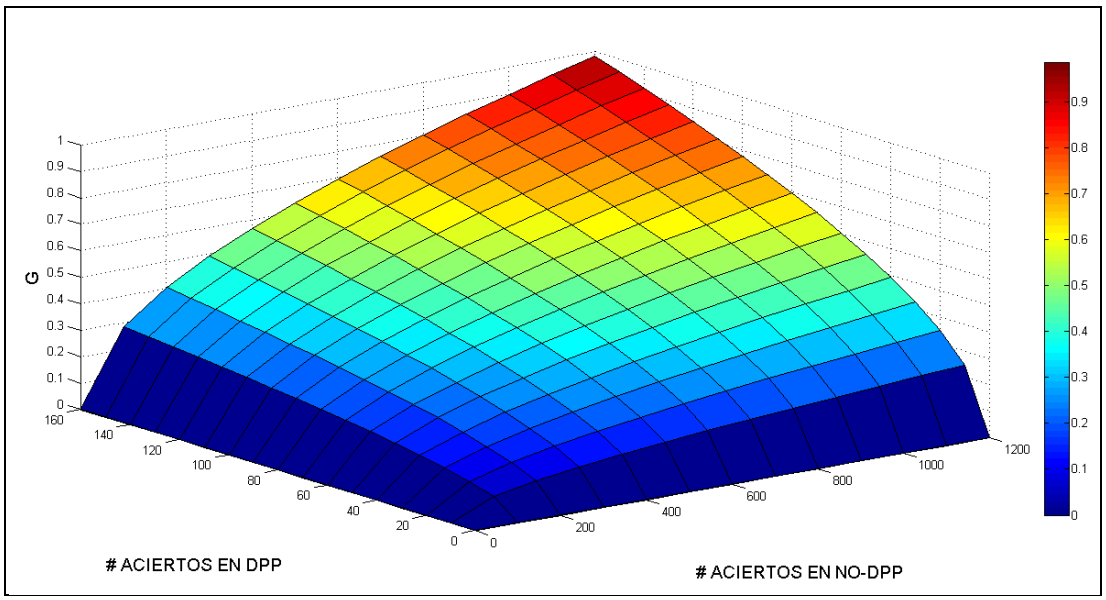


Figura 15: Superficie 3D de la función G según el número de aciertos en las distintas clases de nuestro conjunto de datos.

La media aritmética entre sensibilidad y especificidad ($(SEN + ESP)/2$) también se suele utilizar en la evaluación de modelos, pero en este estudio se prefiere usar la función G ya que es más pesimista.

3. Resultados Analíticos

Durante esta sección se presentarán los resultados analíticos obtenidos al examinar la base de datos que disponemos. Para ello primero se clasificarán las variables independientes según su tipo en categóricas, numéricas discretas y numéricas continuas. Se les aplicará unos test estadísticos y análisis según esta clasificación para demostrar la significancia y relación con la variable dependiente, en este caso la existencia o no de DPP. También se explicará el proceso de preprocesado realizado a los datos antes de entrenar los distintos clasificadores.

Como se ha descrito anteriormente, a partir de las 15 variables independientes descritas, se plantea un problema de clasificación a partir de un conjunto de datos con 1397 instancias, de las que 160 son casos o pacientes enfermas, y las 1237 restantes son controles o pacientes sanas. De las 15 variables, una es numérica continua, siete son numéricas discretas y las siete restantes son de tipo categórico, mostrándose en la Tabla 5 dicha relación.

V. Categóricas	V. Numéricas Discretas	V. Numéricas Continuas
N-Educativo	N-Conviviendo	Edad
Sit-Laboral	EPQN	
Alt-Emocion	N-exp-ini	
Antec-Psi	N-exp-8s	
N-Economic	N-exp-32s	
5-HTT-GC	EPDS	
Sexo	DUKE-sum	

Tabla 5: Variables del dataset según su tipo

Es importante conocer las diferencias que existen entre los tres tipos distintos de variables con las que trabajaremos. Una variable categórica es aquella cuyos valores representan distintas cualidades, características o modalidades, siendo cada una de éstas una categoría. Estas categorías han de ser mutuamente excluyentes, indicando pertenencia o no a un grupo, como por ejemplo tener o no antecedentes psiquiátricos, o haber tenido una situación laboral de desempleada durante el embarazo.

Respecto a las variables numéricas, las variables que disponemos se dividen en dos grupos: discretas y continuas. Las variables numéricas discretas sólo pueden tomar valores dentro de un rango determinado de números enteros, pudiendo considerarse este rango como una escala. Un ejemplo de esto puede ser el número de personas conviviendo en el mismo hogar que la madre, donde en nuestra base de datos va desde 0 a 10. Los cuestionarios de neuroticismo, síntomas depresivos o de apoyo social aportan una escala de puntuaciones que también se consideran discretas.

Finalmente, las variables continuas son aquellas que pueden adquirir cualquier valor a lo largo de un intervalo continuo sin saltos. Normalmente siguen una distribución normal, centrada en una media y con una varianza a ambos lados de ésta. En nuestro estudio se ha considerado la edad de las madres como el único atributo de este tipo, el cual cumple los criterios de normalidad tal y como veremos más adelante.

3.1. Preprocesado de los datos

Una de las etapas más importantes durante el desarrollo de modelos de clasificación es el pretratamiento de los datos antes del entrenamiento de los clasificadores. Según Dorian Pyle en su libro de referencia [57], la preparación de los datos es la clave para resolver un problema y puede suponer la diferencia entre acertar o fallar. Este mismo autor afirma que el objetivo fundamental del preprocesado de datos es manipular y transformar los datos en bruto de modo que el contenido de la información pueda ser más fácilmente entendible y manejable.

Otro objetivo principal del preprocesado es la limpieza de errores contenidos en el conjunto de datos. Estos errores pueden ser de muchos tipos, desde ruido, datos incompletos, duplicados, inconsistencias o valores atípicos. Una inconsistencia puede ser debida a una incorrecta imputación de la información en la base de datos, como por ejemplo un código erróneo o una edad negativa. Un valor atípico (*outlier*) es aquel que se aleja mucho de la media o la mediana. Estadísticamente hablando, si se toma como referencia el valor intercuartil (diferencia entre el primer y tercer cuartil), en un diagrama de caja se considera un valor atípico leve el que se encuentra a 1.5 veces esa distancia de uno de esos cuartiles. Si el valor estuviera a 3 veces esa distancia se consideraría atípico extremo.

Después de una exploración se comprobó que la base de datos a partir de la que se trabaja no presenta ningún error ni valor atípico, excepto en la variable Edad donde existen 13 *outliers*. Estos *outliers* son de tipo leve, y tienen sentido ya que representan a mujeres entre los 18 y 19 años, y los 44 y 46 años, por lo que se opta por no filtrar estas muestras puesto que podrían estar aportando información correcta.

No obstante sí que tenemos valores perdidos *(missing values)* en nuestra base de datos. La falta de respuesta se asocia a diversas causas: a la fatiga del informante, al desconocimiento de la información solicitada, al rechazo de las personas a informar acerca de temas sensibles, así como a problemas asociados a la calidad del marco de muestreo [58]. En la Figura 16 se puede ver un gráfico de barras con el porcentaje de valores perdidos que presenta cada variable, así como en la Tabla 6 se detallan estos porcentajes junto con el número exacto de '*missing values*' por variable independiente en el mismo orden. Cabe destacar que ninguna variable presenta un porcentaje alarmante de falta de datos.

Es necesario pues, seguir una estrategia de imputación de estos valores perdidos, considerándose este proceso parte de la investigación, con el propósito de llegar a conclusiones sustentadas en evidencia empírica sólida durante los test estadísticos y el desarrollo de modelos.

En primer lugar se consideró que las muestras donde existieran más del 30% de variables con valores perdidos debían ser eliminadas del conjunto de datos. Sin embargo ninguna cumplió esta condición, con lo que el *dataset* se mantuvo con sus 1397 registros originales. En segundo lugar se sustituyeron los valores perdidos por la media en el caso de la variable continua Edad, y por la moda en el caso de las demás variables discretas.

Finalmente, las variables categóricas se sustituyeron por variables indicadoras o '*dummy*'. Para ilustrar esta representación tomaremos como ejemplo la variable Nivel Educativo, la cual presenta tres valores distintos (Alto, Medio, Bajo). Se crea una nueva variable por cada uno de las posibles categorías, pudiendo tomar los valores 1 o 0. Esta unidad sólo se activa cuando la variable contiene el valor que le corresponde. En este ejemplo con tres posibles valores, la codificación implica la creación de tres nuevas variables, donde Alto sería asignado a la tupla [1 0 0], Medio a la [0 1 0] y Bajo a la [0 0 1]. Los valores nulos se representaron simplemente mediante la no activación de ninguna de las unidades, siendo la tupla [0 0 0] la correspondiente al ejemplo.

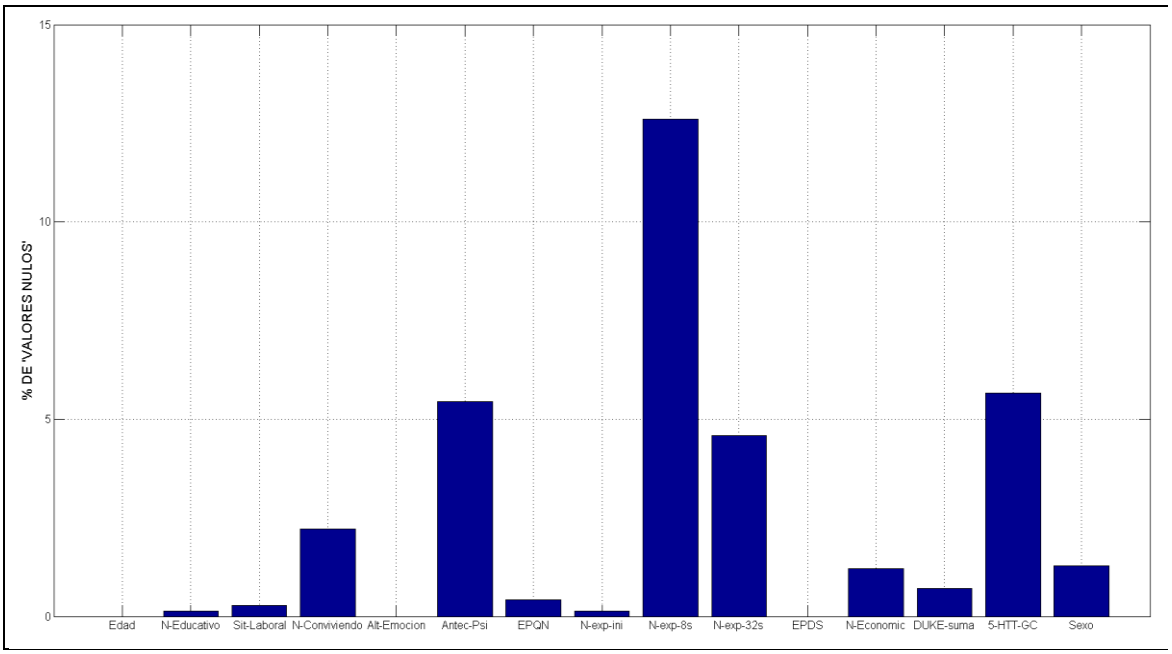


Figura 16: Porcentaje de 'valores perdidos' en cada variable antes de ser imputados.

Variable	# Valores Perdidos	% Valores Perdidos
Edad	0	0.00
N-Educativo	2	0.14
Sit-Laboral	4	0.29
N-Conviviendo	31	2.22
Alt-Emocion	0	0.00
Antec-Psi	76	5.44
EPQN	6	0.43
N-exp-ini	2	0.14
N-exp-8s	176	12.60
N-exp-32s	64	4.58
EPDS	0	0.00
N-Economic	17	1.22
DUKE-suma	10	0.72
5-HTT-GC	79	5.65
Sexo	18	1.29

Tabla 6: Número y porcentaje de valores perdidos en cada variable independiente de la base de datos.

Una vez imputados los valores perdidos, la diferencia de medias y desviaciones estándar en las columnas numéricas antes y después de esta imputación se muestran en la Tabla 7. Es la variable DUKE-suma la que mayor variación sufre al haber hecho este pretratamiento, siendo esto normal ya que su rango de valores es el mayor de todas las variables. Las demás diferencias son prácticamente despreciables. Evidentemente, puesto que hemos sustituido los valores perdidos de las variables discretas con la moda (valor más frecuente), dicha moda no varía. A partir de estos resultados podríamos concluir que no hemos modificado excesivamente el *dataset* tras el preprocesado, y que sigue representando fidedignamente la población de estudio.

Variable	Media (Antes)	Media (Despúes)	Diferencia en la Media	Diferencia. en la Desv. Est.
N_Conviviendo	2.6728	2.6578	0.0149	0.0053
EPQN	3.5219	3.5154	0.0065	0.0046
N-exp-ini	1.0394	1.0379	0.0015	0.0000
N-exp-8s	0.9222	0.8497	0.1225	0.0272
N-exp-32s	0.9985	0.9528	0.0457	0.0088
DUKE-suma	93.8183	93.3185	0.4998	0.0582

Tabla 7: Diferencia en las medias y desviaciones estándar de las variables numéricas antes y después de imputar los valores perdidos.

Todos los análisis exploratorios y modelos de clasificación desarrollados que se mostrarán en los siguientes apartados están hechos a partir de los datos con los valores perdidos ya imputados tal como se ha descrito.

3.2. Análisis exploratorio

A continuación se realizará un análisis exploratorio de los datos diferenciando entre variables categóricas, numéricas discretas y numéricas continuas. Para cada uno de estos tipos se determinarán sus estadísticos, representaciones gráficas, correlaciones y contrastes de hipótesis.

3.2.1. Análisis exploratorio de las Variables Categóricas

Las variables categóricas, como su nombre indica, expresan atributos o características mutuamente excluyentes. En este apartado se ha realizado un análisis exploratorio de las variables de este tipo presentes en nuestro conjunto de datos. En la Figura 17 se muestran los gráficos de tartas de cada variable categórica donde se aprecian sus distintas categorías.

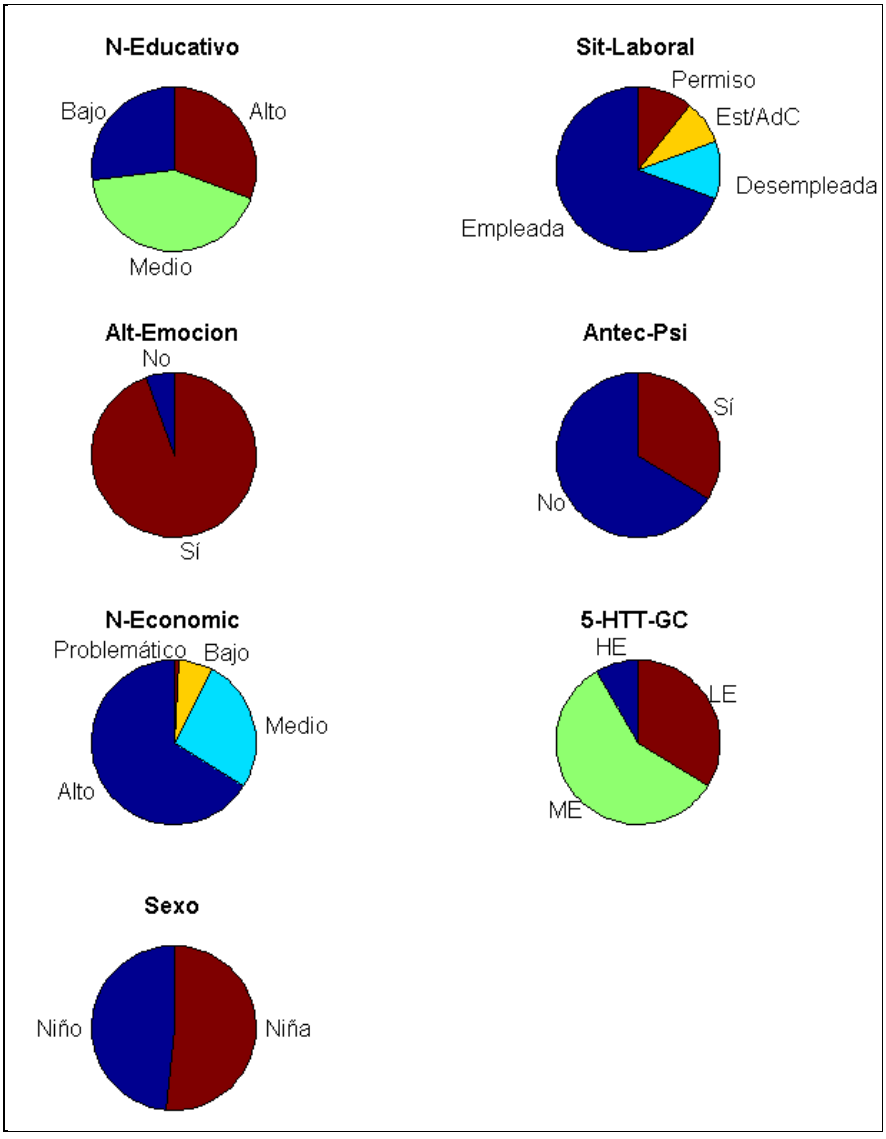


Figura 17: Gráficos de tartas de las variables categóricas.

Se realizó un test estadístico χ^2 (chi-cuadrado) a partir de la tabla de contingencias entre cada variable categórica y la clase, siendo la clase el desarrollo o no de DPP.

El test estadístico χ^2 es utilizado para realizar pruebas de independencia, que nos permite determinar si existe una relación entre dos variables categóricas. Indica si existe o no una relación entre las variables, pero no indica el grado o el tipo de relación; es decir, no muestra el porcentaje de influencia de una variable sobre la otra o la variable que causa la influencia.

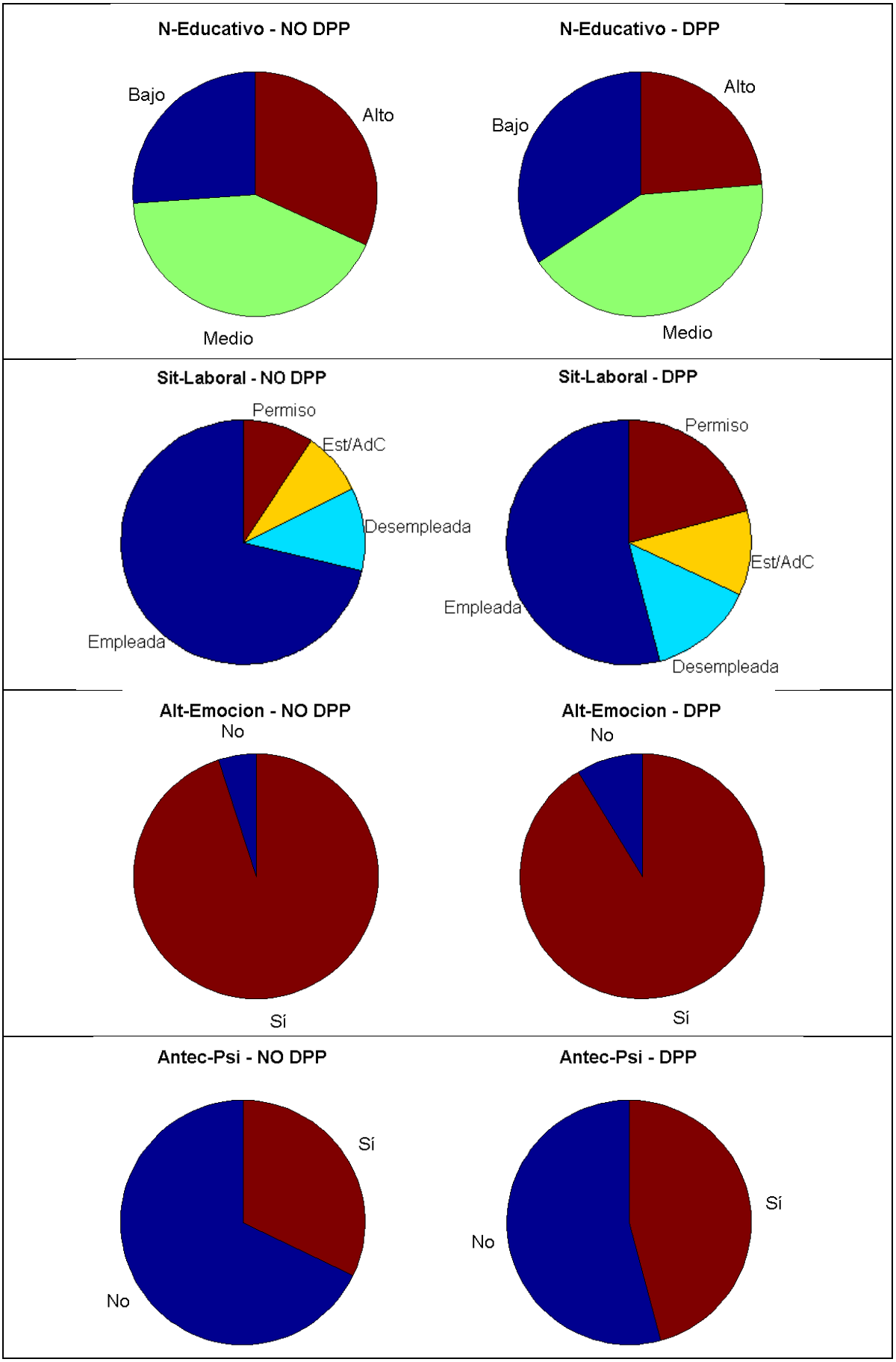
Este test tiene dos parámetros. El primero es *α* (alfa), el cual hace referencia al nivel de confianza que deseamos que tengan los cálculos de la prueba. Es decir, si queremos tener un nivel de confianza del 95%, como es nuestro caso, el valor de α debe ser del 0.05. El segundo parámetro es *k* (grados de libertad) el cual indica el número de categorías independientes en la prueba de independencia.

Para el test estadístico χ^2, la hipótesis nula es que la variable X (categórica) e Y (clase) son independientes, es decir, no están relacionadas. Para poder rechazar esta hipótesis, el valor χ^2 calculado a partir de la tabla de contingencias de cada variable categórica y la clase debe ser superior al $\chi^2_{k,\alpha}$, o lo que es lo mismo, que el valor de *p* de este test sea menor que α. En la Tabla 8 se indica el número de categorías, los grados de libertad (*k*), el valor del test χ^2 y su valor de p para cada una de las variables categóricas. Se aprecia que todas cumplen las condiciones antes mencionadas, siendo sus valores de *p* menores que 0.05. Por lo tanto, podemos rechazar la hipótesis nula en todas ellas y deducir que existe alguna relación o dependencia entre ellas y la clase, siendo estadísticamente significativas. Es por eso que no se prescinde de ninguna de estas variables a la hora de desarrollar los modelos de clasificación.

Variable	# Categorías	*k*	χ^2	*p*
N-Educativo	3	2	6.4421	0.03991
Sit-Laboral	4	3	25.3709	0.00001
Alt-Emocion	2	1	3.8479	0.04981
Antec-Psi	2	1	11.6603	0.00064
N-Economic	4	3	30.4497	0.00000
5-HTT-GC	3	2	6.0423	0.04875
Sexo	2	1	7.8265	0.00515

Tabla 8: Test χ^2 sobre las variables categóricas respecto la clase.

Esta diferencia de proporciones en cada una de las variables según su clase, demostrada en el test χ^2, queda también evidenciada en la Figura 18, donde se muestran los gráficos de tartas de cada variable separando las pacientes que desarrollaron una DPP de las que no. En ella podemos apreciar, por ejemplo, que la proporción de madres que no desarrollaron una DPP es mucho mayor entre las que tuvieron una situación laboral de empleadas durante el embarazo.


N-Educativo - NO DPP
Bajo
Alto
Medio
N-Educativo - DPP
Alto
Bajo
Medio
Sit-Laboral - NO DPP
Permiso
Est/AdC
Desempleada
Empleada
Sit-Laboral - DPP
Permiso
Est/AdC
Empleada
Desempleada
Alt-Emocion - NO DPP
No
Sí
Alt-Emocion - DPP
No
Sí
Antec-Psi - NO DPP
Sí
No
Antec-Psi - DPP
Sí
No

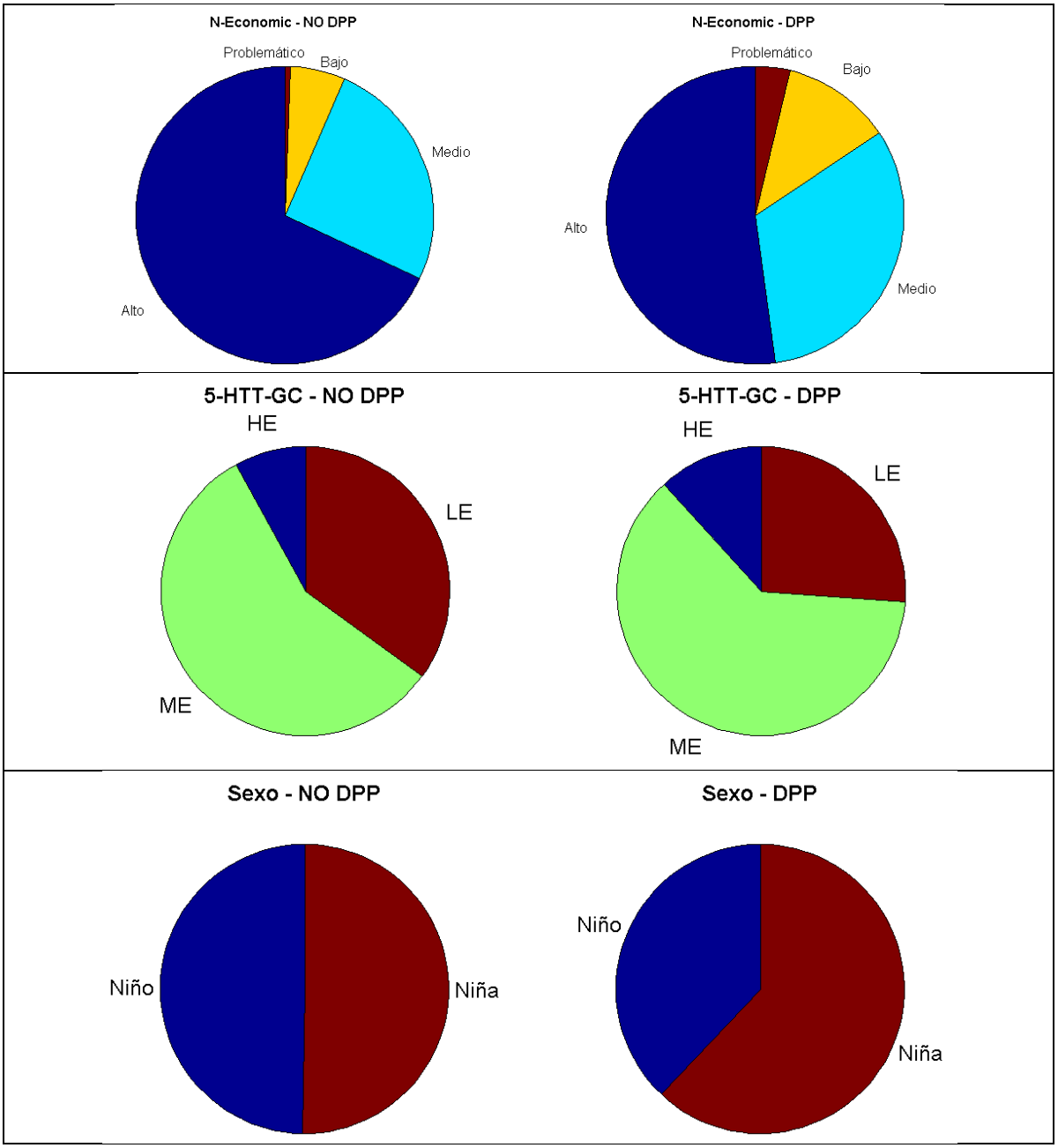


Figura 18: Gráficos de tartas para cada variable categórica separadas según su clase (desarrollo o no de DPP).

Para finalizar este análisis exploratorio, y con el fin de poder encontrar posibles relaciones de variables categóricas entre sí, también se realizó un test χ^2 con todas las posibles combinaciones de ellas. Se encontraron las siguientes relaciones con un nivel de significación estadística del 95%:

- N-Educativo y Sit-Laboral ($k=6$, $\chi^2=80.3022$, $p=0.000000$)
- N-Educativo y N-Economic ($k=6$, $\chi^2=35.3941$, $p=0.000004$)
- Sit-Laboral y N-Economic ($k=4$, $\chi^2=43.9238$, $p=0.000001$)

Estas relaciones se explican fácilmente si se asume que a mayor nivel educativo más posibilidades existen de estar empleada y de disponer de un mejor nivel económico, tal y como sucede en el conjunto de datos.

3.2.2. Análisis exploratorio de las Variables Numéricas Discretas

Las variables numéricas discretas sólo pueden tomar valores dentro de un conjunto numerable. En el siguiente análisis exploratorio de las variables discretas que contiene nuestro conjunto de datos se mostrarán algunas representaciones gráficas, estadísticos básicos y test de independencia respecto la clase.

El polígono de frecuencias es un gráfico que se crea a partir del histograma de frecuencia. Los histogramas emplean columnas verticales para reflejar las frecuencias, mientras que el polígono de frecuencia se construye uniendo los puntos de mayor altura de estas columnas. Por lo tanto un polígono de frecuencias es aquel que se forma a partir de la unión de los distintos puntos medios de las cimas de las columnas del histograma de frecuencia. En la Figura 19 aparecen los polígonos de frecuencias de todas las variables discretas, donde se puede apreciar a simple vista los rangos de estas variables y frecuencia de aparición de los distintos valores entre las 1397 mujeres que participaron en el estudio.

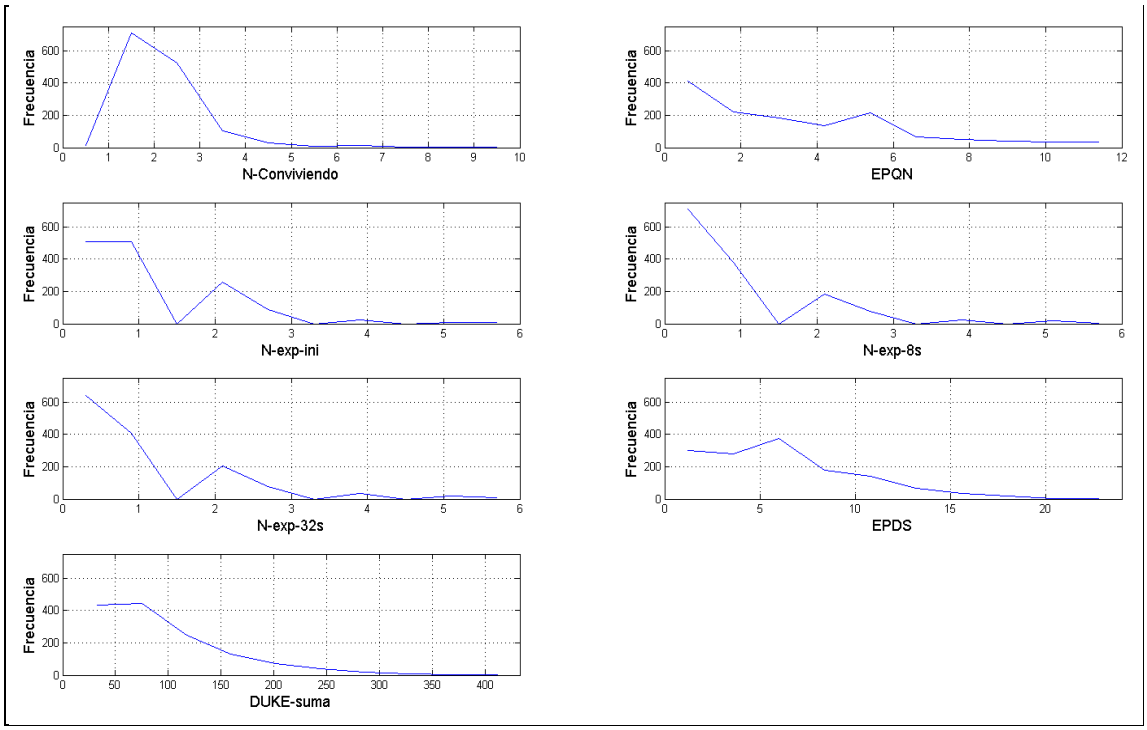


Figura 19: Polígonos de frecuencias de las variables numéricas discretas

La Tabla 9 muestra los estadísticos básicos de las variables discretas, tales como el mínimo, máximo, el intervalo, el número de valores únicos, la mediana y la moda. La variable Duke-suma, la cual mide el apoyo social a la madre, presenta el mayor intervalo de todas, siendo éste de 421. Sin embargo, con 252 valores únicos y una mediana de 77 es fácil concluir que esta escala concentra casi toda la población en valores inferiores a 150, tal como se aprecia en el polígono de frecuencias. Algo parecido sucede con N-

Conviviendo, donde a pesar de existir madres que convivían en un hogar con hasta 10 personas, la mayoría de la población de estudio se concentra en hogares donde residen entre 1 y 4 personas.

Variable	Mín.	Máx.	Intervalo	#Valores Únicos	Mediana	Moda
N-Conviviendo	0	10	10	11	2	2
EPQN	0	12	12	13	3	2
N-exp-ini	0	6	6	7	1	0
N-exp-8s	0	6	6	7	0	0
N-exp-32s	0	6	6	7	1	0
EPDS	0	24	24	23	5	5
DUKE-suma	12	433	421	252	77	24

Tabla 9: Estadísticos básicos sobre las variables numéricas discretas.

Con las variables numéricas discretas también se realizó un test χ^2 para determinar si existen diferencias significativas respecto a la clase. En la Tabla 10 se muestran tanto los grados de libertad (*k*), como el valor χ^2 y *p* del test estadístico. Se puede apreciar que según este test, la presencia o no de DPP (clase) tiene dependencia o relación con todas las variables, excepto N-Conviviendo, con un nivel de confianza del 95% al ser el valor de *p* menor que 0.05. Esto indica que estas variables nos están aportando información para poder clasificar.

Variable	*k*	χ^2	*p*
N-Conviviendo	10	4.7563	0.921163
EPQN	12	126.1019	0.000000
N-exp-ini	6	30.1757	0.000042
N-exp-8s	6	78.3024	0.000000
N-exp-32s	6	160.2975	0.000000
EPDS	22	123.8548	0.000000
DUKE-suma	251	422.8364	0.000000

Tabla 10: Test χ^2 sobre las variables numéricas discretas respecto la clase.

Puesto que en un primer momento la variable N-Conviviendo aparecía como estadísticamente no significativa según el estadístico χ^2, se hicieron pruebas desarrollando un modelo de regresión logística que contemplaba esta variable y otro que prescindía de ella. Se comprobó que las regresiones logísticas que sí incluían esta variable en el entrenamiento presentaban mejores resultados que las que no lo hacían. Por lo tanto, se adoptó la decisión de no prescindir de dicha variable durante el resto del estudio.

3.2.3. Análisis exploratorio de las Variables Numéricas Continuas

En este trabajo se asumió la Edad como la única variable numérica continua del conjunto de datos. Además, esta variable seguía una distribución normal tal y como se detallará más adelante. En la Tabla 11 se muestran sus estadísticos básicos tanto para el conjunto de todas las mujeres que participaron en el estudio, como separando aquellas que no desarrollaron DPP de las que sí lo hicieron. El sesgo negativo indica que existe una ligera cola hacia la izquierda en las distribuciones de los datos tal y como se aprecian en los distintos histogramas de la Figura 21 y la Figura 22. Respecto a la kurtosis, donde se acerca a 3 indica que sigue una distribución normal no picuda [59], tal y como sucede con el conjunto total de los datos de Edad, y con los pertenecientes a madres sin DPP.

	Edad	Edad (NO DPP)	Edad (DPP)
Mín.	18	18	19
Máx.	46	46	43
Rango	28	28	24
Media	32.1274	32.1593	31.8813
Mediana	32	32	32
Moda	32	32	34
Desv. Est.	4.4911	4.4283	4.9582
Sesgo	-0.1551	-0.1772	-0.0006
Kurtosis	3.1933	3.2822	2.6635
Percentil	[29-35]	[29-35]	[29-35]
Kolmogorov-Smirnov (*p*)	0.000124	0.000064	*** 0.458005
Wilcoxon-Mann-Whitney (*p*)	0.000000	-	-

Tabla 11: Estadísticos básicos de la variable Edad en todo el conjunto de datos, así como separando madres con y sin DPP.

También se realizó un test de normalidad de los datos mediante el test de Kolmogorov-Smirnov [60] [61]. Antes de realizar este test se normalizaron los datos a unidades tipificadas, normalización también conocida como *z-score*, donde a cada valor (x) se le resta la media (μ) y se divide entre la desviación estándar (σ) de la variable a la que pertenece, tal y como indica la ecuación de la Figura 20.

$$z = \frac{x - \mu}{\sigma}$$

Figura 20: Ecuación básica de normalización a unidades tipificadas o *z-score*

La hipótesis nula del test de Kolmogorov-Smirnov indica que los datos no siguen una distribución normal. Para rechazar esta hipótesis nula con un nivel de confianza del 95% el valor de *p* en este test ha de ser menor que 0.05. De acuerdo a los resultados obtenidos, podemos decir que el conjunto de todos los datos de Edad siguen una distribución normal, cumpliéndose también esto para los datos que pertenecen sólo a madres sin DPP. Sin embargo, puesto que el valor de *p* es mayor que 0.05 en el caso de las mujeres que sí fueron diagnosticadas de DPP, podemos decir que este subconjunto de datos no cumple la condición de normalidad.

Resumiendo los resultados de normalidad de los datos, Edad y Edad(NO DPP) siguen una distribución normal. No ocurre lo mismo con Edad(DPP), seguramente por el pequeño número de muestras que se disponen de esta clase. Esto se puede apreciar en la Figura 21 *y* Figura 22 donde aparecen los histogramas de esta variable para todas las pacientes, y también separándolos entre las que desarrollaron DPP y las que no.

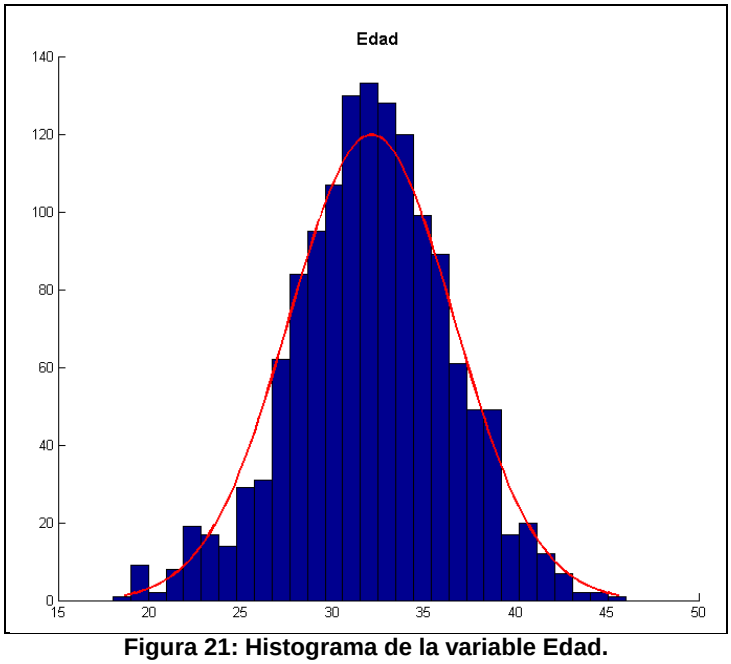


Figura 21: Histograma de la variable Edad.

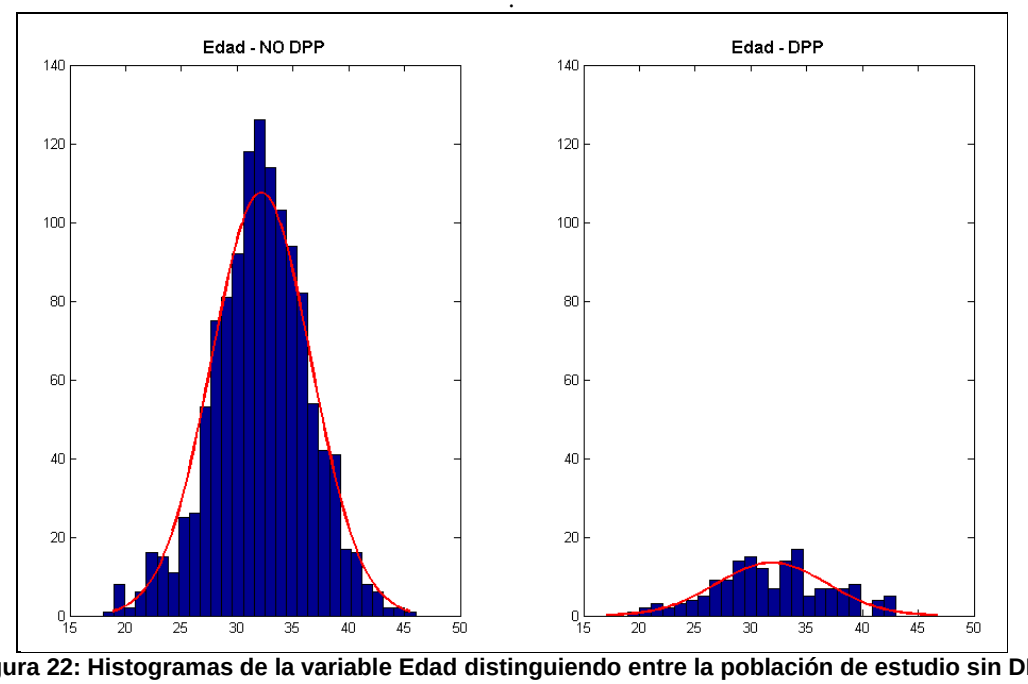


Figura 22: Histogramas de la variable Edad distinguiendo entre la población de estudio sin DPP y la que finalmente sí la desarrolló.

Para finalizar este análisis exploratorio, también se realizó un test de igualdad de medianas respecto la clase. El test empleado fue la prueba de suma de rangos de Wilcoxon-Mann-Whitney [62] [63]. La hipótesis nula de este test indica que dos variables X (Edad-NO DPP) e Y (Edad-DPP) provienen de la misma distribución. Para rechazar

esta hipótesis nula con un nivel de confianza del 95%, el valor de *p* del test ha de ser inferior a 0.05.

El resultado obtenido en nuestro caso es un valor de p de 0, tal y como se indica en la Tabla 11. Por lo tanto podríamos rechazar la hipótesis nula, y afirmar que Edad(NO DPP) y Edad(DPP) no provienen de la misma distribución de datos. Podemos confirmar de esta manera las conclusiones a las que llegamos con el test de Kolmogorov-Smirnov, el cual indicaba que la edad de las mujeres que no desarrollaron DPP seguía una distribución normal, mientras que no sucedía lo mismo con las que sí la sufrieron. Esto contrasta con el diagrama de caja o *boxplot* de la Figura 23, donde a simple vista no se aprecia dicha diferencia de distribuciones, siendo debido en parte a que las medianas son iguales. Los valores atípicos u *outliers* podrían estar afectando en este caso. Cabe destacar que tenemos 12 *outliers* en el grupo de las mujeres sin DPP, y sólo 1 en el de las que desarrollaron la enfermedad. En conclusión, según estos resultados donde se indica que existen diferencias en las distribuciones de datos entre las distintas clases, Edad sería una variable con poder discriminatorio a la hora de incluirla en los modelos a entrenar.

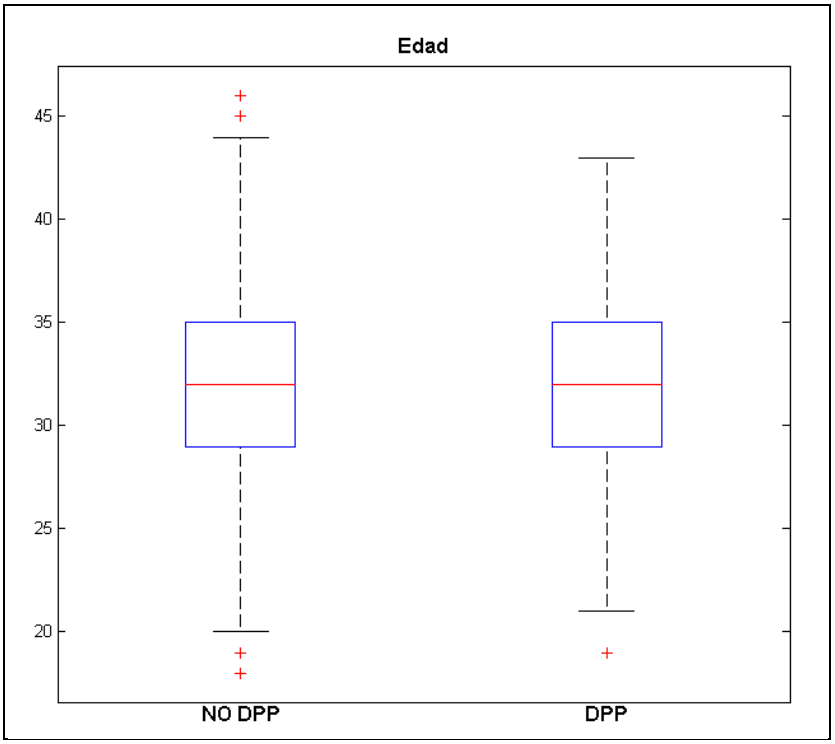


Figura 23: Boxplot o diagrama de caja de la variable Edad.

4. Resultados Experimentales

En esta sección se presentarán los resultados experimentales obtenidos durante el entrenamiento de los diferentes clasificadores probados. El orden de presentación comienza con el modelo Naïve Bayes, siendo sus resultados los que se consideran como los mínimos a superar por los demás clasificadores que aparecen en este estudio. A continuación se describen los experimentos y resultados con la Regresión Logística y con los modelos de tipo SVM (*Support Vector Machines*). Los últimos experimentos que se detallan son los realizados con RNA (Redes Neuronales Artificiales), siendo este tipo de modelo el que mejor rendimiento presenta. Finalmente se muestra una visión global de los resultados experimentales, indicando los modelos seleccionados para su implementación en dispositivos móviles y su rendimiento tras el reentrenamiento con todos los datos.

En secciones anteriores se detalló el proceso de división de los datos en tres conjuntos: entrenamiento, validación y evaluación. En ellos se mantienen las proporciones entre casos (DPP) y controles (NO DPP). Además, la diferencia en la media de edad de estos tres conjuntos no supera el valor de 0.1 y la mayor diferencia en la desviación típica de edad es de 0.25 entre el conjunto de entrenamiento y el de evaluación. Damos así por buena esta división de los datos. Cabe destacar que en los modelos donde no existe validación de hiperparámetros (Naïve Bayes y Regresión Logística) los conjuntos de entrenamiento y validación se fusionan en uno solo.

También es necesario recordar que el objetivo de este trabajo es desarrollar dos modelos diferentes. Uno para ser utilizado en la aplicación móvil para personal clínico al que llamamos modelo MÉDICO y que contempla todas las variables del conjunto de datos descritas previamente. Y otro que será usado en la aplicación destinada a madres al que llamamos modelo PACIENTE, del que se excluyen cinco variables: EPQN, DUKE-suma, N-Exp-8s, N-Exp-32s y 5-HTT-GC. Es por eso que en este último modelo caben esperar menores prestaciones.

Para todos los modelos se indica los resultados de clasificación en todos los conjuntos de datos utilizados en sus respectivos experimentos (entrenamiento, evaluación y validación si es necesario). Estos resultados son la Sensibilidad (SEN), Especificidad (ESP), Valor Predictivo Positivo (VPP), Valor Predictivo Negativo (VPN), Razón de Verosimilitud Positiva (RV+), Razón de Verosimilitud Negativa (RV-) y Matriz de Confusión (MC). También se indica el acierto o *Accuracy* (ACC) junto con su Intervalo de Confianza (IC) al 5% de significación estadística. Además, se incluye el valor de G, siendo el resultado más relevante de este estudio ya que nos da una medida de relación entre sensibilidad y especificidad, siendo $G = \sqrt{SEN \cdot ESP}$. Esta última medida en el conjunto de datos de evaluación será utilizada como criterio de valoración de los distintos modelos MÉDICO y PACIENTE. En nuestro caso es importante obtener alta sensibilidad sin perder excesiva especificidad en el modelo MÉDICO, mientras que en el modelo PACIENTE sería recomendable obtener una alta especificidad con una razonable sensibilidad.

Para los modelos donde es necesario un umbral de decisión en su salida para discriminar entre madres sin depresión postparto (NO DPP) y con depresión postparto (DPP), también se muestran las curvas ROC, junto con el valor del área bajo la curva (AUC) obtenido. Estos modelos son la Regresión Logística y las RNA, siendo en ambos la salida un valor comprendido entre 0 y 1.

Además, para estos modelos con umbral de decisión o punto de corte, se han creado unos gráficos que muestran los individuos separados según su clase, junto con el valor que nos ha dado el clasificador como salida. El mejor umbral posible también aparece representado. Más adelante se explicará con mayor detalle este tipo de gráfico.

4.1. Naïve Bayes

En este apartado se presentarán los resultados experimentales obtenidos mediante Naïve Bayes, tanto en el modelo MÉDICO como en el PACIENTE.

En la Tabla 12 se muestran todas las estadísticas de clasificación mediante Naïve Bayes con el modelo MÉDICO. Se consiguen unos nada despreciables valores de G de 0.68 y 0.67 en los conjuntos de entrenamiento y evaluación respectivamente, teniendo en cuenta la simplicidad de este tipo de modelo. Al estar estos dos resultados muy equilibrados en ambos conjuntos podemos afirmar que no existe sobreajuste con los datos de entrenamiento, y que por lo tanto generaliza correctamente. De las demás medidas destaca la alta especificidad que presenta (0.90). Además, su valor de VPP de 0.4 es el mayor de todos los clasificadores que veremos más adelante. Esto significa que cuando el modelo indica presencia de DPP, no está incluyendo en este grupo a demasiadas personas sanas por error. Sin embargo, al tener una sensibilidad tan baja (0.50), hace que la medida G caiga en comparación a los demás clasificadores.

NAÏVE BAYES MODELO MÉDICO	Entrenamiento	Evaluación
G	0.685	0.672
SEN	0.523	0.500
ESP	0.897	0.903
VPP	0.396	0.400
VPN	0.936	0.933
ACC+IC	0.854 [0.834-0.875]	0.857 [0.816-0.898]
RV+	5.080	5.146
RV-	0.531	0.554
MC	$\begin{pmatrix} 67 & 102 \\ 61 & 888 \end{pmatrix}$	$\begin{pmatrix} 16 & 24 \\ 16 & 223 \end{pmatrix}$

Tabla 12: Resultados de clasificación con Naïve Bayes. Modelo MÉDICO.

Por otro lado, la Tabla 13 muestra todas las estadísticas de clasificación mediante Naïve Bayes con el modelo PACIENTE. En este caso los valores de G son de 0.5 y 0.45 en los conjuntos de entrenamiento y evaluación respectivamente. Estos resultados se pueden considerar como bastante pobres, y son debidos a su baja sensibilidad, llegando al 0.22 en el conjunto de evaluación, a pesar de la elevada especificidad de 0.93. Por lo tanto nos indica que este clasificador está balanceando en exceso los datos hacia el grupo de las personas sanas. El desbalanceo en las clases y el inferior número de datos predictores al eliminar las 5 variables correspondientes al modelo PACIENTE parece afectar al rendimiento de Naïve Bayes.

NAÏVE BAYES MODELO PACIENTE	Entrenamiento	Evaluación
G	0.506	0.450
SEN	0.281	0.219
ESP	0.911	0.927
VPP	0.290	0.280
VPN	0.907	0.902
ACC+IC	0.839 [0.817-0.861]	0.846[0.804-0.888]
RV+	3.164	3.002
RV-	0.789	0.843
MC	$\begin{pmatrix} 36 & 88 \\ 92 & 902 \end{pmatrix}$	$\begin{pmatrix} 7 & 18 \\ 25 & 229 \end{pmatrix}$

Tabla 13: Resultados de clasificación con Naïve Bayes. Modelo PACIENTE.

Estos resultados se presentan como punto de partida. Como ya se ha mencionado, también se realizó una experimentación con los mismos datos de entrenamiento y evaluación mediante Árboles de Decisión y KNN. El mejor valor de G en el conjunto de evaluación con el modelo MÉDICO fue de 0.57 con los Árboles, y de 0.46 con KNN tomando los 3 vecinos más próximos. Con el modelo PACIENTE estos resultados tampoco mejoraron en comparación con Naïve Bayes.

A continuación se presentarán los resultados de los modelos que sí mejoraron el rendimiento de Naïve Bayes, los cuales fueron la Regresión Logística, *Support Vector Machines* (SVM) y las Redes Neuronales Artificiales (RNA).

4.2. Regresión Logística

En este apartado se presentarán los resultados experimentales obtenidos mediante la Regresión Logística, tanto en el modelo MÉDICO como en el PACIENTE.

Para ambos modelos, se estimó el vector de coeficientes β_i para una regresión lineal generalizada en las respuestas de Y (clase) sobre los predictores de X (variables), usando una distribución binomial. En una distribución binomial, Y puede ser un vector binario indicando 'éxito' o 'fracaso' en cada observación [64]. En nuestro caso, Y es un vector con las respuestas observadas, es decir, si una madre desarrolló o no DPP.

Tras ajustar los coeficientes β_i de la regresión logística con los datos de entrenamiento y obtener los valores de la salida tanto para datos de entrenamiento como para los de evaluación, se puede ajustar el umbral óptimo para poder clasificar. Valores superiores o iguales a dicho umbral indicarán pertenencia a la clase positiva (DPP), e inferiores a la negativa (NO DPP). Para ello se seleccionó de entre todo el rango de valores de salida aquel que maximizaba la relación entre sensibilidad y especificidad con el conjunto de datos de entrenamiento. En nuestro caso, como se ha mencionado anteriormente, utilizamos la función G ($G = \sqrt{SEN \cdot ESP}$) para representar esta relación.

El punto de corte o umbral óptimo en la salida del modelo MÉDICO resultó ser de 0.17, obteniéndose un valor de G de 0.77 con los datos de entrenamiento y de 0.72 en los de evaluación. Esta forma de selección del umbral tiene el inconveniente de impedir tratar la salida como una probabilidad, ya que se desplaza el punto de corte de 0.5 a uno seleccionado por nosotros. Sin embargo los resultados en la clasificación mejoran considerablemente con esta opción.

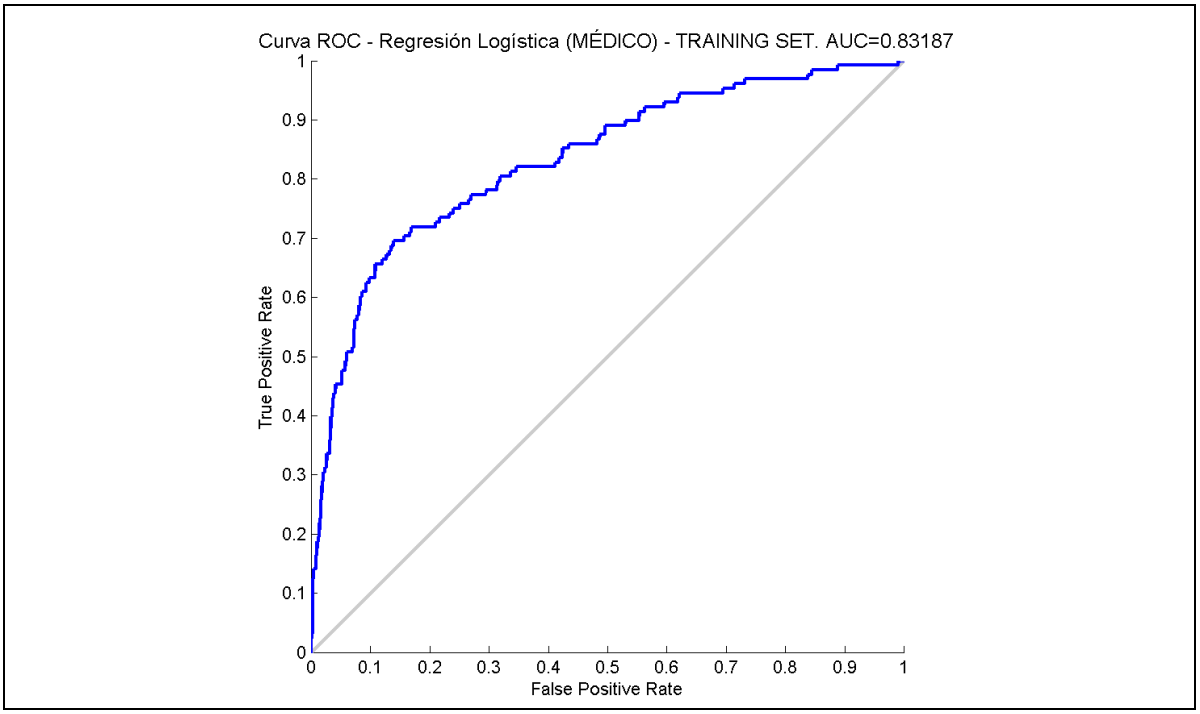
Respecto a los parámetros β_i de la regresión, los correspondientes a las variables N-exp-ini, DUKE-suma y la unidad correspondiente al Sexo femenino del bebé resultaron ser de -0.0062, 0.0062 y 0.0018 respectivamente. Al ser ligeramente inferiores a 0.01 se probó con una nueva regresión logística eliminando estas variables. Sin embargo, los resultados de G y AUC fueron inferiores a los conseguidos antes de eliminarlas, con lo que se optó por mantener dichas variables en el modelo.

En la Tabla 14 se muestran todas las estadísticas de clasificación de la Regresión Logística con el modelo MÉDICO. Las medidas que peores resultados presentan son la sensibilidad y VPP. Es decir, la tasa de aciertos detectando casos (DPP) no supera el 0.63 con los datos de evaluación, y además, puesto que tenemos un VPP bajo (0.31), si clasifica una madre como enferma, la probabilidad de que realmente lo esté es baja. Sin embargo, en cuanto a las medidas de especificidad y VPN, relacionadas con la tasas de aciertos en controles (NO DPP), se obtienen resultados bastante buenos. La especificidad de 0.82 en datos de evaluación indica que detecta a gran parte de las madres sanas, y el VPN de 0.94 muestra que si la prueba da un resultado negativo, existe una alta probabilidad de que realmente sea así. Hay que tener presente en el análisis de estos resultados el desbalanceo existente entre las clases.

REG. LOGÍSTICA MODELO MÉDICO	Entrenamiento	Evaluación
G	0.774	0.715
AUC	0.832	0.826
SEN	0.695	0.625
ESP	0.862	0.818
VPP	0.394	0.308
VPN	0.956	0.944
ACC+IC	0.843 [0.821-0.864]	0.796 [0.748-0.843]
RV+	5.025	3.431
RV-	0.354	0.459
MC	$\begin{pmatrix} 89 & 137 \\ 39 & 853 \end{pmatrix}$	$\begin{pmatrix} 20 & 45 \\ 12 & 202 \end{pmatrix}$

Tabla 14: Resultados de clasificación con Regresión Logística. Modelo MÉDICO.

En la Figura 24 se muestran las curvas ROC obtenidas con la Regresión Logística para el modelo MÉDICO, tanto con los datos de entrenamiento como con los de evaluación. Se puede apreciar que ambas AUC son prácticamente iguales, con valores cercanos al 0.83, con lo que se podría afirmar que este modelo no está sobreajustado a los datos de entrenamiento y que sus resultados son moderadamente fiables.



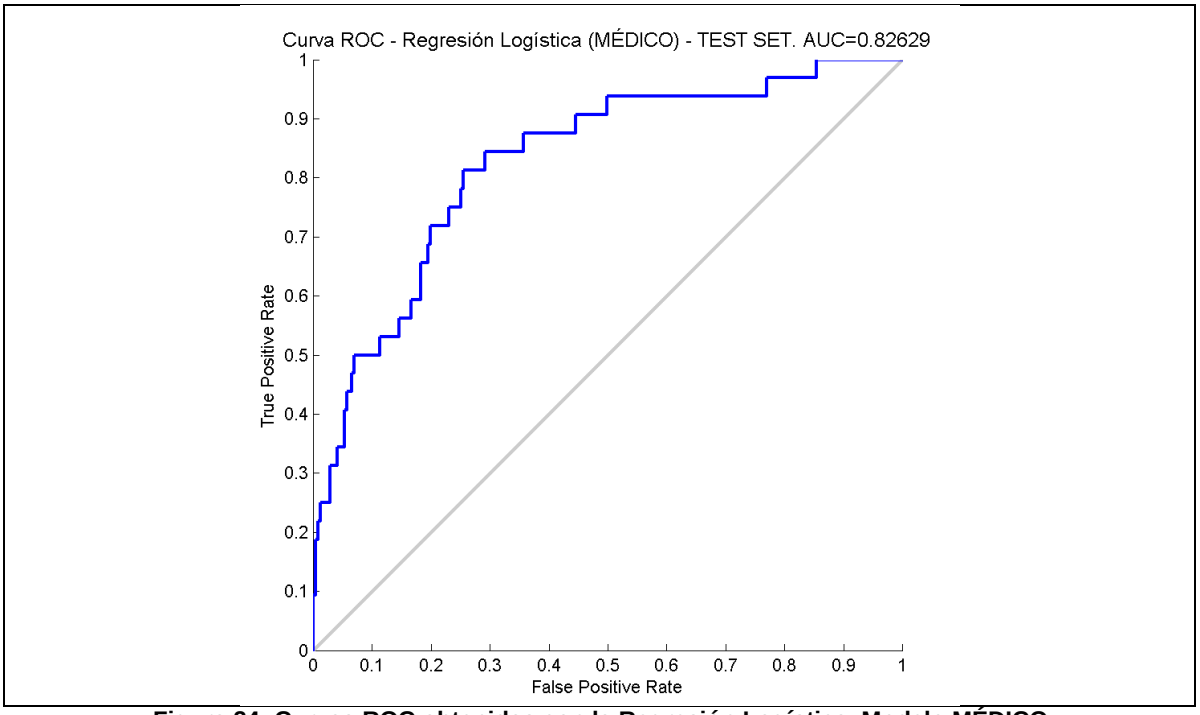


Figura 24: Curvas ROC obtenidas con la Regresión Logística. Modelo MÉDICO.

Como ya se ha mencionado, la Regresión Logística necesita un umbral de decisión o punto de corte en los valores de su salida para discriminar casos y controles. La Figura 25 muestra la salida de la Regresión Logística y este umbral de decisión frente a la clase para el modelo MÉDICO, tanto con los datos de entrenamiento como con los de evaluación. En este gráfico se separa a las participantes del estudio según la clase, siendo los puntos verdes los correspondientes a madres sanas (NO DPP), y los rojos a enfermas (DPP). El eje de ordenadas (y) indica el valor de la salida del clasificador para cada paciente. El umbral de decisión óptimo, calculado a partir de los resultados de la salida del modelo con los datos de entrenamiento es de cerca de 0.17 y se muestra en la gráfica mediante una línea horizontal azul. Todos los individuos que obtienen un valor superior o igual a este umbral en la regresión logística los consideramos como enfermos, mientras que los que no lo superen se consideran sanos. Es decir, los aciertos del clasificador se representan mediante los puntos verdes por debajo del umbral (Verdaderos Negativos) y los puntos rojos por encima del umbral (Verdaderos Positivos). Los fallos se representan mediante los puntos verdes por encima del umbral (Falsos Positivos) y los puntos rojos por debajo del umbral (Falsos Negativos).

Los valores elevados de especificidad y VPN avalan la afirmación de que este clasificador funciona bastante bien acertando en la detección de mujeres sanas. No sucede así con los casos (DPP), que abarcan prácticamente todo el espectro de la salida y cuyos valores de sensibilidad y VPP son bastante más bajos que los anteriores de especificidad y VPN. El desbalanceo en las clases es una de las razones por las que esto sucede.

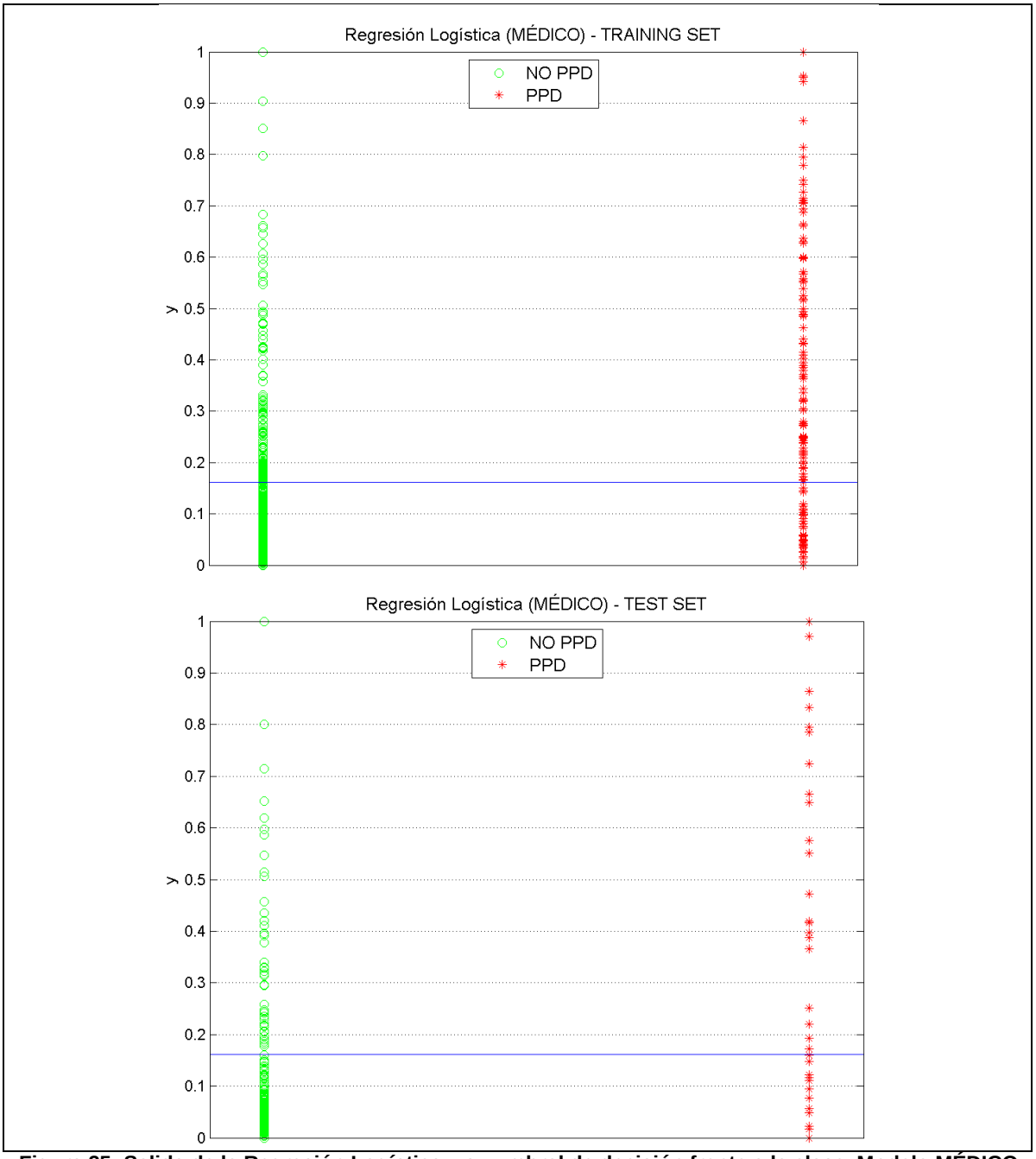


Figura 25: Salida de la Regresión Logística y su umbral de decisión frente a la clase. Modelo MÉDICO.

Respecto a la experimentación de la Regresión Logística con el modelo PACIENTE, también se estimó el vector de coeficientes β_i de la misma forma que se hizo con el modelo MÉDICO. Todos los coeficientes β_i resultaron ser superiores a 0.01, con lo que todas las variables independientes se consideraron relevantes.

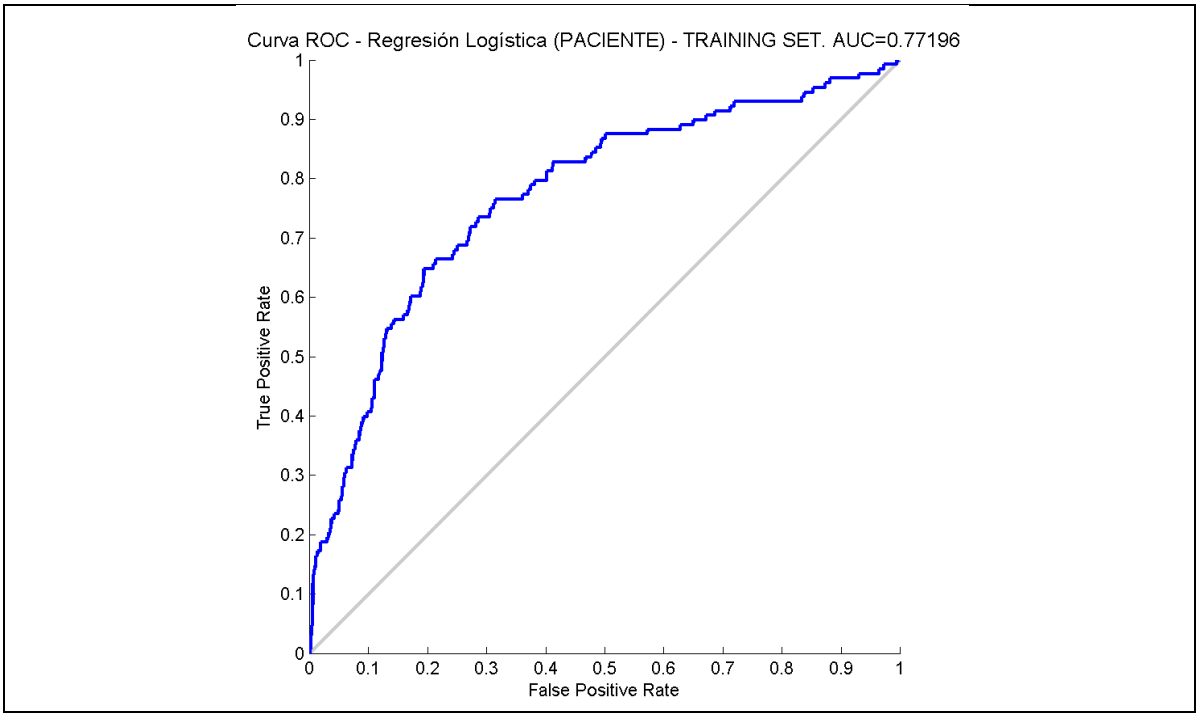
Se seleccionó el umbral óptimo de esta Regresión Logística a través de la salida con los datos de entrenamiento siendo su valor cercano a 0.1. Al igual que en el modelo MÉDICO, valores superiores o iguales a dicho umbral indicarán pertenencia a la clase positiva (DPP), e inferiores a la negativa (NO DPP). El valor de G obtenido con este umbral fue de 0.73 con los datos de entrenamiento y de 0.71 en los de evaluación.

En la Tabla 15 se muestran todas las estadísticas de clasificación de la Regresión Logística con el modelo PACIENTE. Aquí nos sucede lo contrario que en el modelo MÉDICO, teniendo mejores tasas de sensibilidad que de especificidad. La sensibilidad es de 0.75 en el conjunto de evaluación, y la especificidad de 0.67. Ambas medidas son bastante buenas teniendo en cuenta el desbalanceo de clases y que este modelo (PACIENTE) contempla menos variables que el anterior (MÉDICO). Además, un resultado de 0.95 en VPN indica la alta probabilidad de acierto si la prueba ofrece un resultado negativo.

REG. LOGÍSTICA MODELO PACIENTE	Entrenamiento	Evaluación
G	0.725	0.706
AUC	0.772	0.748
SEN	0.766	0.750
ESP	0.686	0.664
VPP	0.240	0.224
VPN	0.958	0.953
ACC+IC	0.695 [0.668-0.722]	0.674 [0.619-0.729]
RV+	2.437	2.232
RV-	0.342	0.377
MC	$\begin{pmatrix} 98 & 311 \\ 30 & 679 \end{pmatrix}$	$\begin{pmatrix} 24 & 83 \\ 8 & 164 \end{pmatrix}$

Tabla 15: Resultados de clasificación con Regresión Logística. Modelo PACIENTE.

En la Figura 26 se muestran las curvas ROC obtenidas con la Regresión Logística para el modelo PACIENTE, tanto con los datos de entrenamiento como con los de evaluación. Al igual que en el modelo MÉDICO, ambas AUC son también muy parecidas entre sí (0.77 y 0.75 respectivamente). Por ello podemos volver a afirmar en este caso que el modelo no está sobreajustado a los datos de entrenamiento, siendo sus resultados moderadamente fiables.



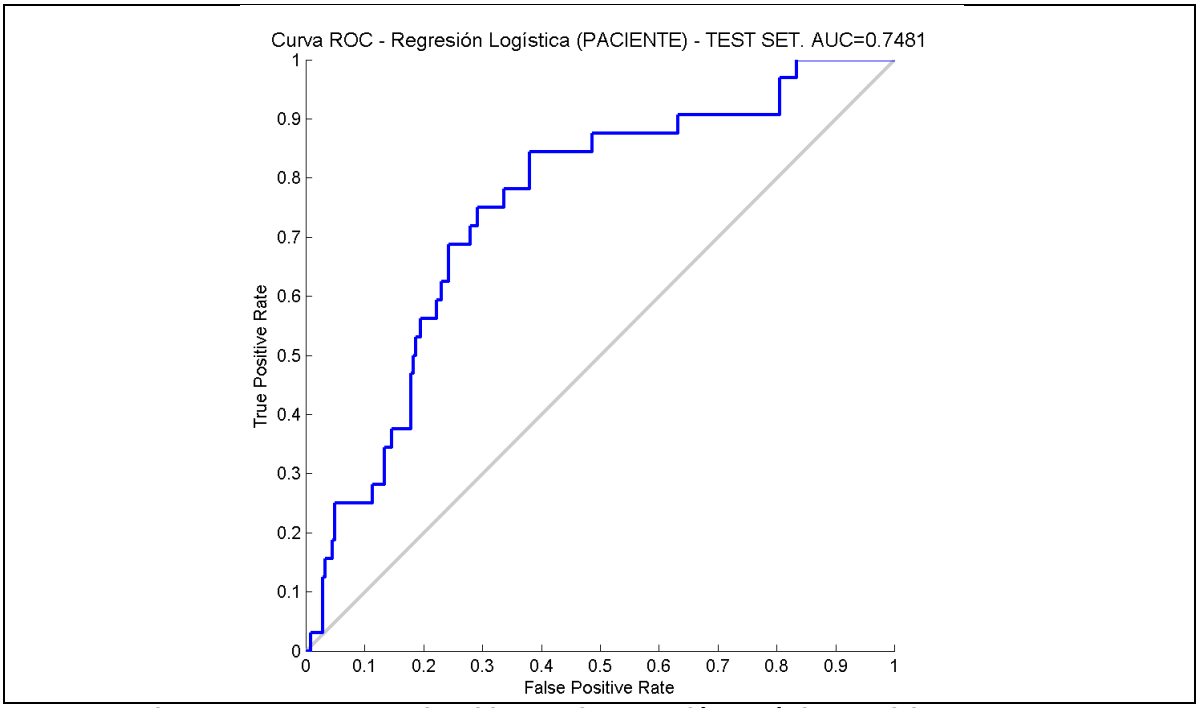


Figura 26: Curvas ROC obtenidas con la Regresión Logística. Modelo PACIENTE.

La Figura 27 muestra la salida de la Regresión Logística y el umbral de decisión frente a la clase para el modelo PACIENTE, tanto con los datos de entrenamiento como con los de evaluación.

Como se puede apreciar, el umbral óptimo en este caso se sitúa cerca de 0.1, siendo este valor bastante pequeño. Esto provoca que, a pesar de tener una sensibilidad relativamente alta para este problema (0.77 y 0.75 en entrenamiento y test respectivamente), la especificidad cae (0.69 y 0.66) considerando como enfermas a muchas madres que no lo son. Esto se ve fácilmente en la gran densidad de puntos verdes que aparecen por encima del umbral (línea horizontal azul). También se aprecia que pocos casos representados con puntos rojos (PPD) alcanzan el valor máximo de la salida, lo que podría indicar un pequeño problema en la calibración de esta Regresión Logística además de estar ante un problema de difícil separación lineal.

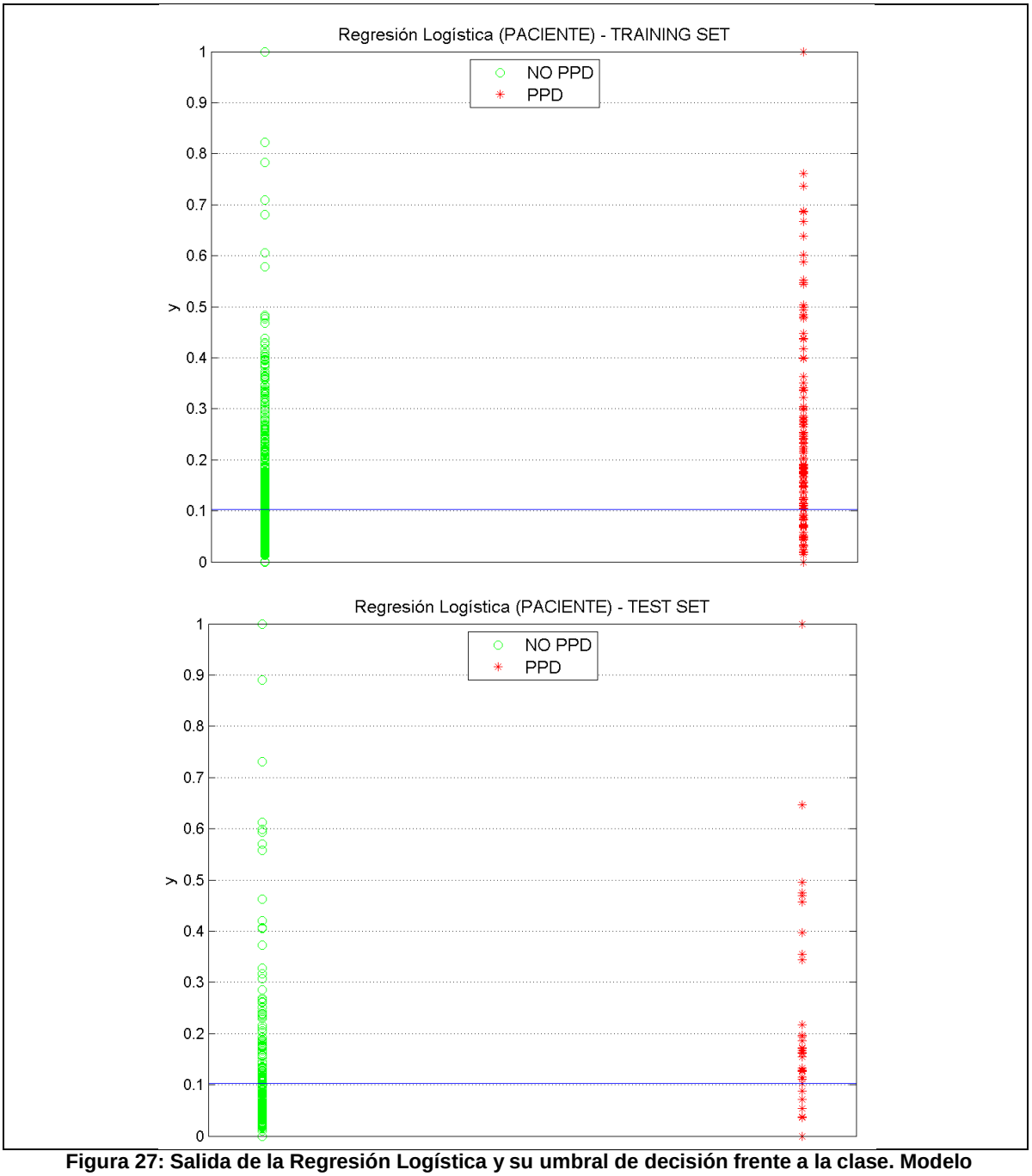


Figura 27: Salida de la Regresión Logística y su umbral de decisión frente a la clase. Modelo PACIENTE.

Como más adelante podremos ver, los mejores resultados de G con los distintos clasificadores de tipo PACIENTE fueron los obtenidos mediante esta Regresión Logística y la RNA. Por esta razón se considerará qué comportamiento ideal necesitamos para la aplicación móvil destinada a las madres y elegir entre estos dos modelos aquel que mejor se ajuste a nuestras necesidades.

.

4.3. Support Vector Machines (SVM)

En este apartado se presentarán los resultados experimentales obtenidos mediante Support Vector Machines (SVM), tanto en el modelo MÉDICO como en el PACIENTE.

Como se ha descrito anteriormente, en SVM se clasifica mediante el hiperplano que maximiza el margen entre dos clases en los datos de entrenamiento. Los vectores que definen este hiperplano, seleccionados de entre los predictores o variables independientes, son los llamados '*support vectors*'.

Con el fin de permitir cierta flexibilidad, los SVM manejan una variable C que controla la compensación entre errores de entrenamiento y los márgenes rígidos, creando así un margen blando (*soft margin*) que permita algunos errores en la clasificación a la vez que los penaliza [49]. Así pues el hiperparámetro C representa el compromiso entre el tamaño del margen y el número de errores de clasificación. Para todos los kernels y sus variaciones de hiperparámetros que se describen a continuación, se hizo una búsqueda en C en el rango [0.6, 1.0] en pasos de 0.1.

Tal y como muestra la Tabla 16, se experimentó con tres tipos de *kernels* distintos: polinomial homogéneo, función de base radial Gaussiana, y perceptrón. La experimentación con kernels polinómicos consistió en hacer una búsqueda en el orden del polinomio (n) desde 1 hasta 7. En los kernels de base radial Gaussiana esta búsqueda fue en el hiperparámetro gamma (γ) desde 3.0 hasta 15.0 en intervalos de 0.1. En los kernels de tipo perceptrón, la búsqueda se realizó en todas las combinaciones posibles de alfa (α) y beta (β), en los rango [0.1, 2.0] y [-1.0, -0.1] respectivamente, también en intervalos de 0.1. En total, si contamos todas las combinaciones de hiperparámetros descritas teniendo también en cuenta la penalización del margen C, se probaron 1635 parametrizaciones con SVM, tanto en la versión MÉDICO como en la PACIENTE.

Kernel	Función	Parámetros testeados
Polinómico	$k(x_i, x_j) = (x_i \cdot x_j)^n$	$C \in [0.6, 1.0] \quad n \in [1, 7]$
F. base radial Gaussiana	$k(x_i, x_j) = e^{\gamma \|x_i - x_j\|^2}$	$C \in [0.6, 1.0] \quad \gamma \in [3.0, 15.0]$
Perceptrón	$k(x_i, x_j) = \tanh(\alpha x_i x_j + \beta)$	$C \in [0.6, 1.0]$ $\alpha \in [0.1, 2.0] \quad \beta \in [-1.0, -0.1]$

Tabla 16: Kernels y rangos de sus hiperparámetros probados con SVM.

La efectividad del SVM depende de la selección del *kernel*, los parámetros del *kernel* y el parámetro de penalización del margen C. Debido a estos hiperparámetros a ajustar, se utilizó el conjunto de entrenamiento para entrenar el modelo, y el conjunto de validación para seleccionar aquellos hiperparámetros que mejores resultados presentaban en su clasificación. Finalmente, también se calculan los resultados en el conjunto de evaluación, el cual no ha participado en ningún momento ni en la fase de entrenamiento ni en la de validación. Con SVM no existe selección de ningún umbral de decisión en su salida, ya que la clasificación es binaria al etiquetar cada caso a un lado u otro del hiperplano creado.

En el modelo MÉDICO, los mejores resultados se obtuvieron con un valor de C de 0.7 y un *kernel* de función de base radial con un valor de γ de 9.3, mostrándose en la Tabla 17 todas las estadísticas de clasificación. Con un valor de G de 0.74 en el conjunto de evaluación podemos decir que mejoramos ligeramente el rendimiento obtenido en Naïve Bayes y la Regresión Logística, donde este mismo valor de G estaba en 0.67 y 0.71 respectivamente. Globalmente, la sensibilidad es cercana a 0.70 y la especificidad a 0.80. Valores de VPN de 0.95 nos indican que un resultado negativo en la prueba representa una alta probabilidad de que realmente sea así.

SVM MODELO MÉDICO	Entrenamiento	Validación	Evaluación
G	0.764	0.712	0.737
SEN	0.721	0.625	0.688
ESP	0.810	0.812	0.789
VPP	0.329	0.300	0.297
VPN	0.957	0.944	0.951
ACC+IC	0.80 [0.77-0.83]	0.79[0.74-0.85]	0.78[0.73-0.83]
RV+	3.790	3.321	3.266
RV-	0.344	0.462	0.396
MC	$\begin{pmatrix} 75 & 153 \\ 29 & 651 \end{pmatrix}$	$\begin{pmatrix} 15 & 35 \\ 9 & 151 \end{pmatrix}$	$\begin{pmatrix} 22 & 52 \\ 10 & 195 \end{pmatrix}$

Tabla 17: Resultados de clasificación con SVM. Modelo MÉDICO.

Respecto al modelo PACIENTE, los mejores resultados se obtuvieron con un valor de C de 0.6 y un *kernel* de función de base radial con un valor de γ de 14.6, mostrándose en la Tabla 17 todas las estadísticas de clasificación. Con un valor de G de 0.67 en el conjunto de evaluación vemos como se mejora sustancialmente sobre Naïve Bayes (0.45), pero no sobre la Regresión Logística (0.71). Se obtiene una sensibilidad de 0.60 y especificidad de 0.75 en este mismo conjunto de datos. También se consiguen valores de VPN muy similares a los del modelo MÉDICO anterior, sin embargo, con resultados inferiores a 0.3 en VPP es difícil afirmar que una nueva paciente está enferma si el modelo dice que es así. Como se había supuesto inicialmente, la falta de los datos de las 5 variables excluídas en el modelo PACIENTE afecta significativamente el rendimiento del SVM, ya que con el modelo MÉDICO se consiguieron resultados mucho mejores.

SVM MODELO PACIENTE	Entrenamiento	Validación	Evaluación
G	0.715	0.726	0.669
SEN	0.663	0.667	0.594
ESP	0.771	0.790	0.753
VPP	0.273	0.291	0.237
VPN	0.947	0.948	0.935
ACC+IC	0.76 [0.73-0.79]	0.78[0.72-0.83]	0.74[0.68-0.79]
RV+	2.899	3.179	2.404
RV-	0.436	0.422	0.539
MC	$\begin{pmatrix} 69 & 184 \\ 35 & 620 \end{pmatrix}$	$\begin{pmatrix} 16 & 39 \\ 8 & 147 \end{pmatrix}$	$\begin{pmatrix} 19 & 61 \\ 13 & 186 \end{pmatrix}$

Tabla 18: Resultados de clasificación con SVM. Modelo PACIENTE.

Como hemos podido ver, SVM presenta hasta ahora los mejores resultados obtenidos en el modelo MÉDICO, siendo superado por la Regresión Logística en el modelo PACIENTE. Sin embargo, las RNA mejoran todavía más la clasificación con nuestra base de datos, tal y como se detalla en el siguiente apartado.

4.4. Redes Neuronales Artificiales (RNA)

En este apartado se presentarán los resultados experimentales obtenidos mediante Redes Neuronales Artificiales (RNA), tanto en el modelo MÉDICO como en el PACIENTE. Los hiperparámetros a determinar en una RNA son referentes a su topología (número de capas ocultas y número de unidades por cada capa). Los parámetros que nos ajustará el algoritmo de entrenamiento serán los valores de la matriz de pesos W y vector de *bias* b (sesgo). Según la regla de la pirámide geométrica para la creación de la topología de una RNA propuesta por *Masters* [65], siendo n el número de entradas y m el número de salidas, el número de unidades con una única capa oculta vendría determinado por la expresión $H = \sqrt{n \cdot m}$. En caso de una topología con dos capas ocultas, siendo $r = \sqrt[3]{n/m}$, el número de unidades en la primera capa vendría determinado por $H_1 = m \cdot r^2$, y el de la segunda por $H_2 = m \cdot r$. Teniendo en cuenta estas directrices el número de unidades de la capa siguiente siempre es inferior al de la anterior formando una especie de pirámide en su topología.

En nuestro caso, la codificación utilizada para el modelo MÉDICO hace que el número de entradas sea de 26 (debido a la codificación de variables categóricas) y sólo una salida. De esta forma una topología de RNA con una sola capa oculta debería tener 5 unidades, y otra RNA con dos capas ocultas, 9 unidades en la primera capa y 3 unidades en la segunda. En el caso del modelo PACIENTE este número de entradas se reduce a 19. Tomando lo mencionado como referencia, se optó por hacer un barrido de topologías con dos, una y ninguna capa oculta para ambos modelos (MÉDICO y PACIENTE). En el caso de dos capas ocultas, se experimentó con un rango H_1 que fue en orden decreciente desde 15 hasta 2, mientras que H_2 fue también en orden decreciente desde una unidad menos que H_1 hasta 2. El mismo rango desde 15 a 2 se utiliza para las RNA con una única capa oculta. Como prueba adicional en este experimento, se probaron RNAs sin capas ocultas, las cuales son equivalentes a una Regresión Logística.

Los resultados de entrenamiento de una RNA tienen la particularidad de que dependen de la inicialización aleatoria de la matriz de pesos de sus unidades, ya que esto hace que el resultado final converja en un error mínimo local o uno global. Es por esto que se ha de repetir el proceso de inicialización aleatoria de los pesos (W) y *bias* (b) para intentar caer en un punto inicial que lleve a un mínimo global. En este estudio se optó por repetir esta inicialización 500 veces por cada topología probada.

La Tabla 19 muestra un resumen de las combinaciones de topologías con las que se experimentó.

Capas Ocultas	N° Inicializaciones	H_1	H_2	Total Inicializaciones
2	500	[15, 2]	$[(H_1 - 1), 2]$	49000
1	500	[15, 2]	-	7000
0	500	-	-	500
				56500

Tabla 19: Combinaciones de topologías de redes neuronales de la experimentación.

En esta tabla se puede ver el número de capas ocultas, el rango de unidades en cada capa y el número de inicializaciones aleatorias de W y b en cada topología. Si contamos todas las combinaciones posibles, obtenemos un total de 56500, lo que provocó que el experimento completo tardase del orden de 7 días en finalizar.

La mejor arquitectura de red y sus parámetros fueron seleccionados empíricamente usando el conjunto de validación, para más adelante ser evaluada con el conjunto de test. El algoritmo de entrenamiento de las RNA utilizado fue el *Levenberg-Marquardt backpropagation*, el cual actualiza los valores de W y b según la optimización de *Levenberg-Marquardt* por descenso de gradiente [66]. Para evitar el sobreajuste se establecieron como condiciones de parada del aprendizaje alcanzar un máximo de 100 *epochs* o que la diferencia del gradiente entre dos *epochs* consecutivos fuese inferior a 0.0001. Cabe mencionar que el número de *epochs* se define en este contexto como el número de iteraciones sobre un conjunto de datos para entrenar una RNA. Las funciones de activación de todas las unidades de las RNA fue la tangente hiperbólica sigmoidea, también conocida como *tansig* [67]. Hay que tener en cuenta que esta función devuelve un número en el rango [-1, 1], pero en la capa de salida de nuestras redes neuronales se hace un escalado final de ella a [0, 1].

Una vez entrenada cada RNA, y puesto que la salida es una función *tansig*, para cada red neuronal se calculó el mejor umbral o punto de corte para poder clasificar entre casos (DPP) y controles (NO DPP), tal y como se hizo con la Regresión Logística. Es decir, se seleccionó aquel umbral que maximizaba la relación entre sensibilidad y especificidad (función G) con el conjunto de entrenamiento.

En el modelo MÉDICO, los mejores resultados se obtuvieron con una RNA de dos capas ocultas, y cuya topología era de [26-4-2-1], tal y como se ilustra en la Figura 30.

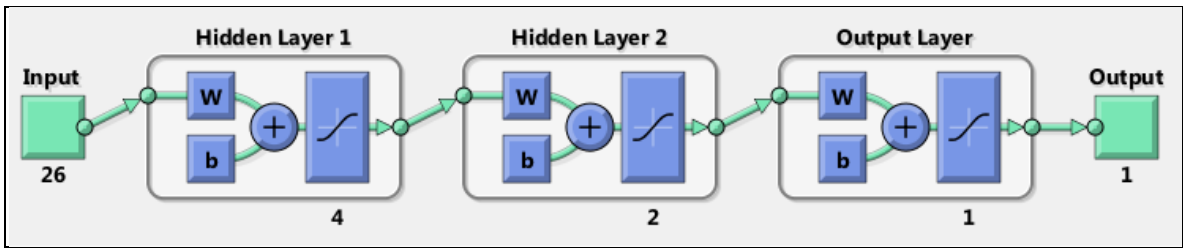


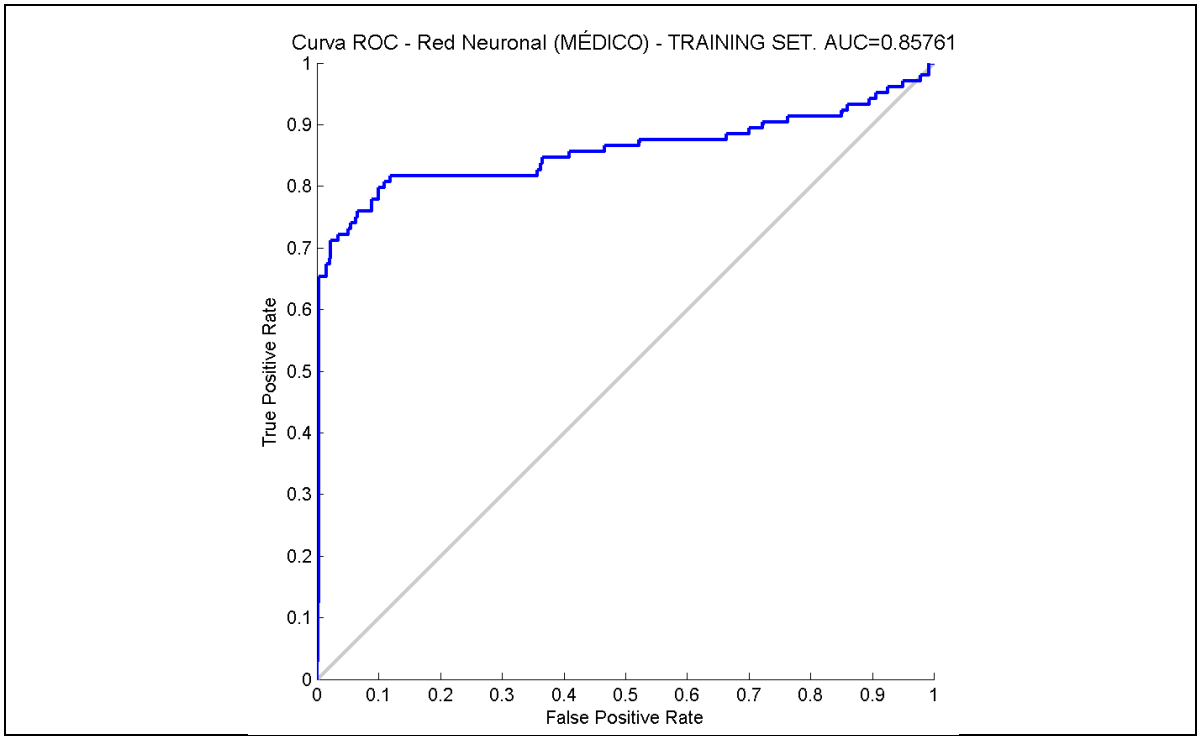
Figura 28: Topología de la mejor red neuronal encontrada. Modelo MÉDICO.

En la Tabla 20 se muestran todas las estadísticas de clasificación con esta RNA para el modelo MÉDICO. Se aprecia que con un valor de G de 0.80 en evaluación se obtiene el mejor clasificador de todo el estudio. Esto es gracias a la sensibilidad de 0.78 y especificidad de 0.82 que presenta, balance muy alto y equilibrado en ambos valores.

RED NEURONAL MODELO MÉDICO	Entrenamiento	Validación	Evaluación
G	0.849	0.804	0.801
AUC	0.858	0.805	0.850
SEN	0.817	0.792	0.781
ESP	0.882	0.817	0.822
VPP	0.472	0.358	0.362
VPN	0.974	0.968	0.967
ACC+IC	0.87 [0.85-0.90]	0.81[0.76-0.87]	0.82[0.77-0.86]
RV+	6.917	4.331	4.386
RV-	0.207	0.255	0.266
MC	$\begin{pmatrix} 85 & 95 \\ 19 & 709 \end{pmatrix}$	$\begin{pmatrix} 19 & 34 \\ 5 & 152 \end{pmatrix}$	$\begin{pmatrix} 25 & 44 \\ 7 & 203 \end{pmatrix}$

Tabla 20: Resultados de clasificación con la mejor red neuronal encontrada. Modelo MÉDICO.

En la Figura 29 se muestran las curvas ROC obtenidas con dicha RNA para el modelo MÉDICO, en los datos de entrenamiento, validación y evaluación. Las AUC en entrenamiento y evaluación son de 0.85. Al ser sus valores prácticamente iguales podemos afirmar que este modelo no está sobreajustado a los datos de entrenamiento, siendo sus resultados de clasificación bastante fiables. La diferencia con la AUC en validación (0.80) seguramente es debida a que este conjunto de datos es más pequeño que los otros dos, propiciando que la forma de la curva ROC para este conjunto sea la más irregular de las tres.



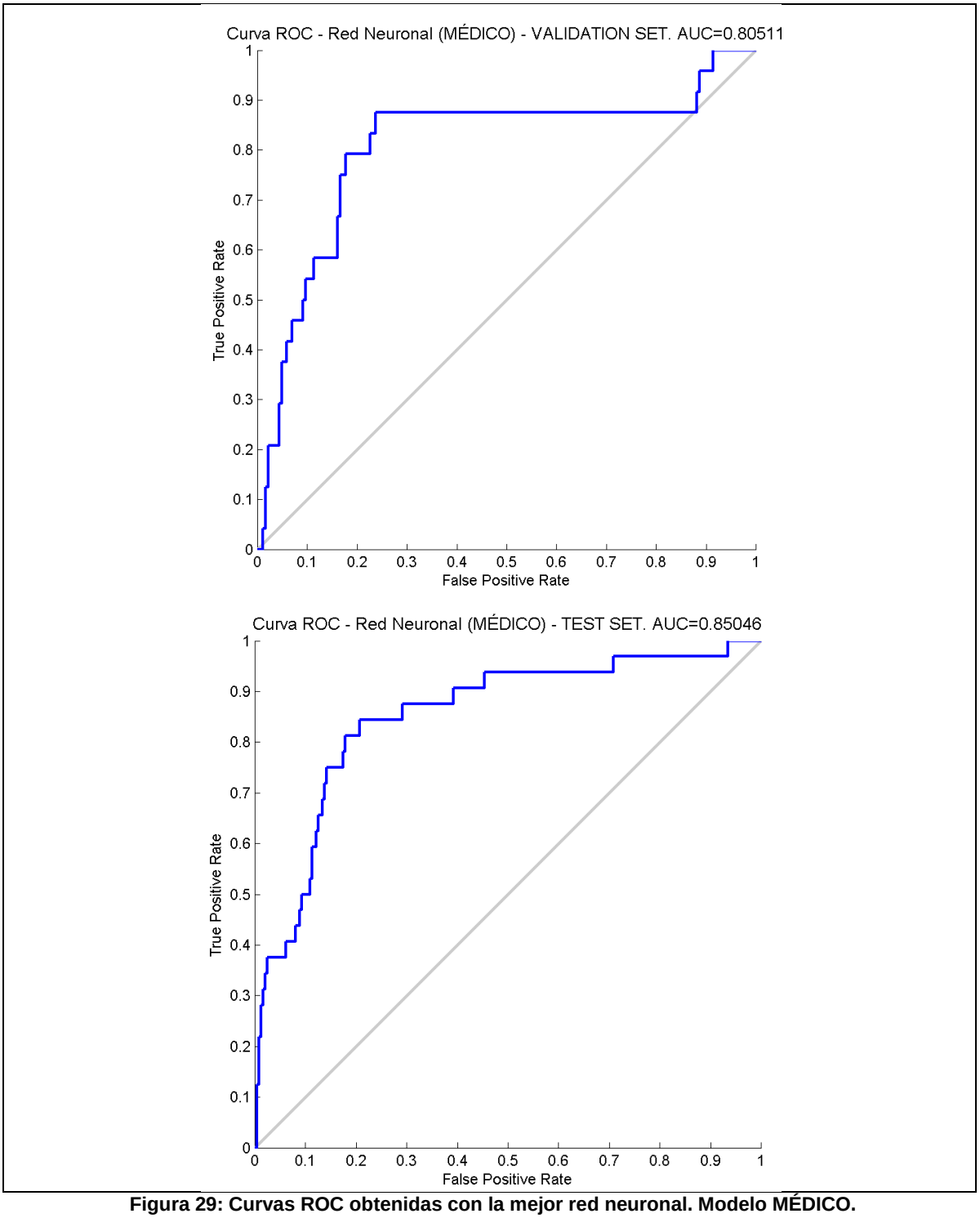


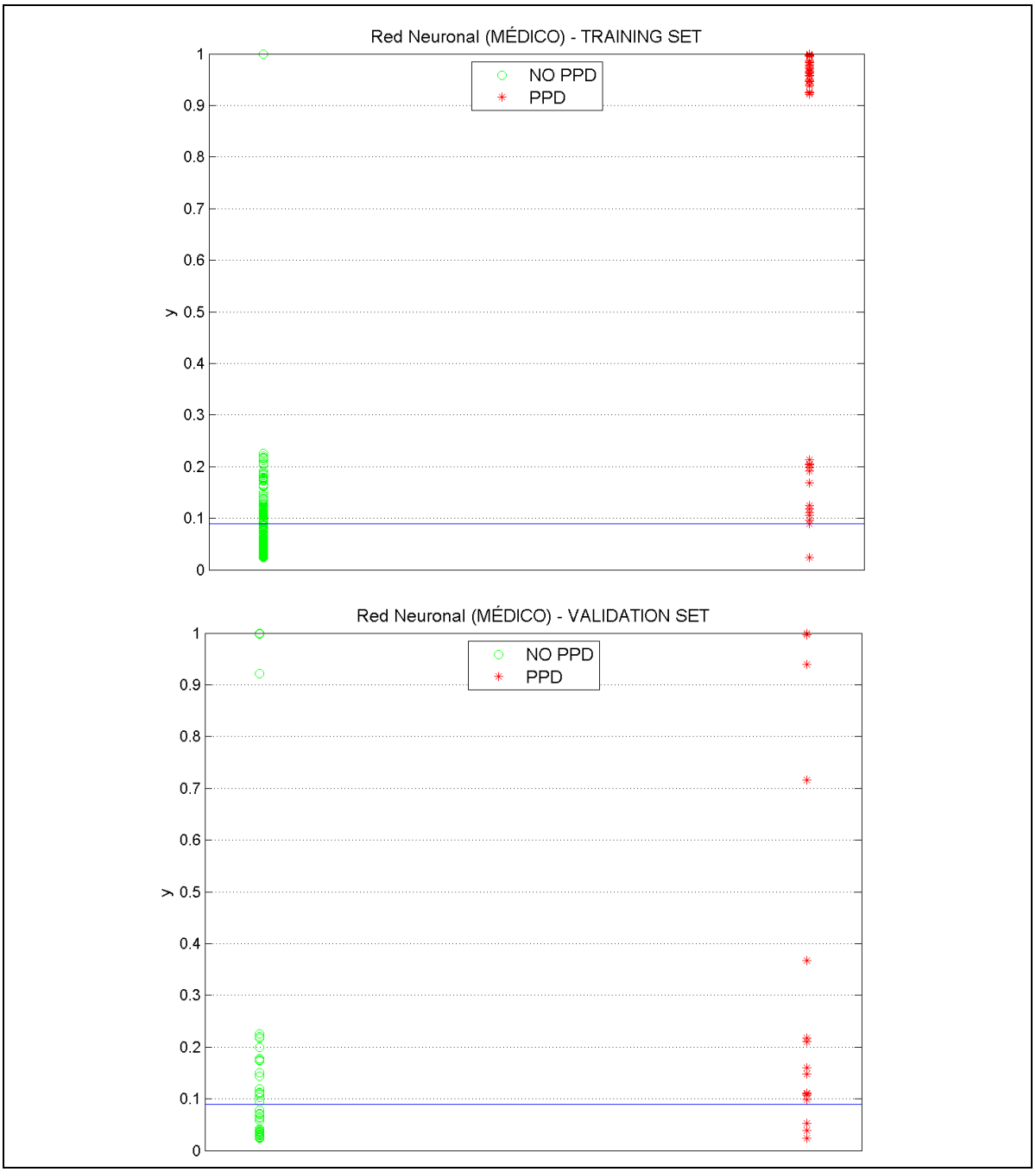
Figura 29: Curvas ROC obtenidas con la mejor red neuronal. Modelo MÉDICO.

La Figura 30 muestra la salida de la RNA y el umbral de decisión frente a la clase para el modelo MÉDICO, con los datos de entrenamiento, validación y evaluación. El tipo de gráfico es exactamente el mismo descrito en la Regresión Logística, donde el eje de ordenadas (y) representa la salida.

El umbral óptimo en este caso es de 0.09. Aunque este valor es bastante pequeño se aprecia como la mayoría de madres enfermas (DPP) se sitúan por encima de este umbral lo que provoca la alta sensibilidad mencionada anteriormente (0.82 y 0.78 en

entrenamiento y evaluación). Además, y como prueba de una buena calibración, la mayoría de la población sana (NO DPP) queda por debajo del umbral, superando en pocos casos el 0.5 en los tres conjuntos de datos. Esto hace que obtengamos también buenos resultados en la especificidad (0.88 y 0.82 en entrenamiento y evaluación).

Por todas las razones indicadas hasta ahora respecto a la mejor RNA obtenida para el modelo MÉDICO, se concluye que es el clasificador que mejor generaliza nuestro problema con todas las variables independientes. Por lo tanto, se reentrenará con todos los datos y migrará a la correspondiente aplicación móvil, explicándose los detalles en la siguiente sección de este estudio.



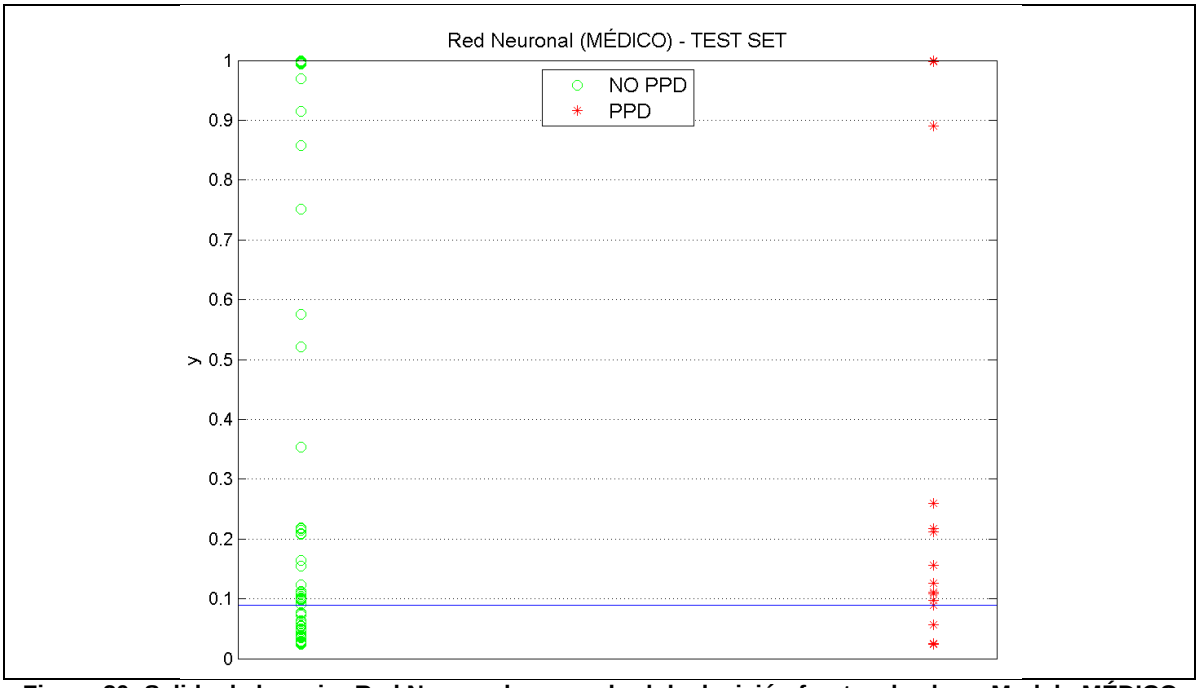


Figura 30: Salida de la mejor Red Neuronal y su umbral de decisión frente a la clase. Modelo MÉDICO.

Respecto a las RNA con las que se experimentó en el modelo PACIENTE, se siguieron los mismos criterios descritos anteriormente. Es decir, las mismas combinaciones de topologías, número de inicializaciones, algoritmo de entrenamiento y condiciones de parada junto con funciones de activación de las unidades que en el modelo MÉDICO. Hay que tener en cuenta que en este caso, y debido a la codificación de las variables categóricas, tenemos un total de 19 entradas. En el modelo PACIENTE, los mejores resultados se obtuvieron con una RNA de una capa oculta, y cuya topología era de [19-2-1], tal y como se ilustra en la Figura 31.

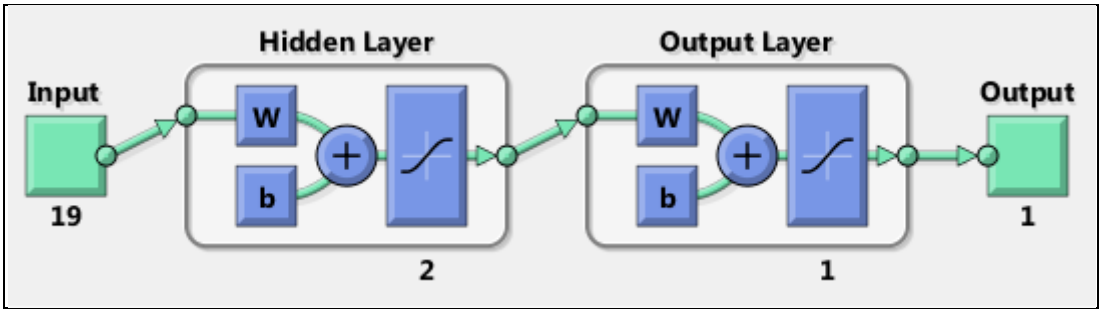


Figura 31: Topología de la mejor red neuronal encontrada. Modelo PACIENTE.

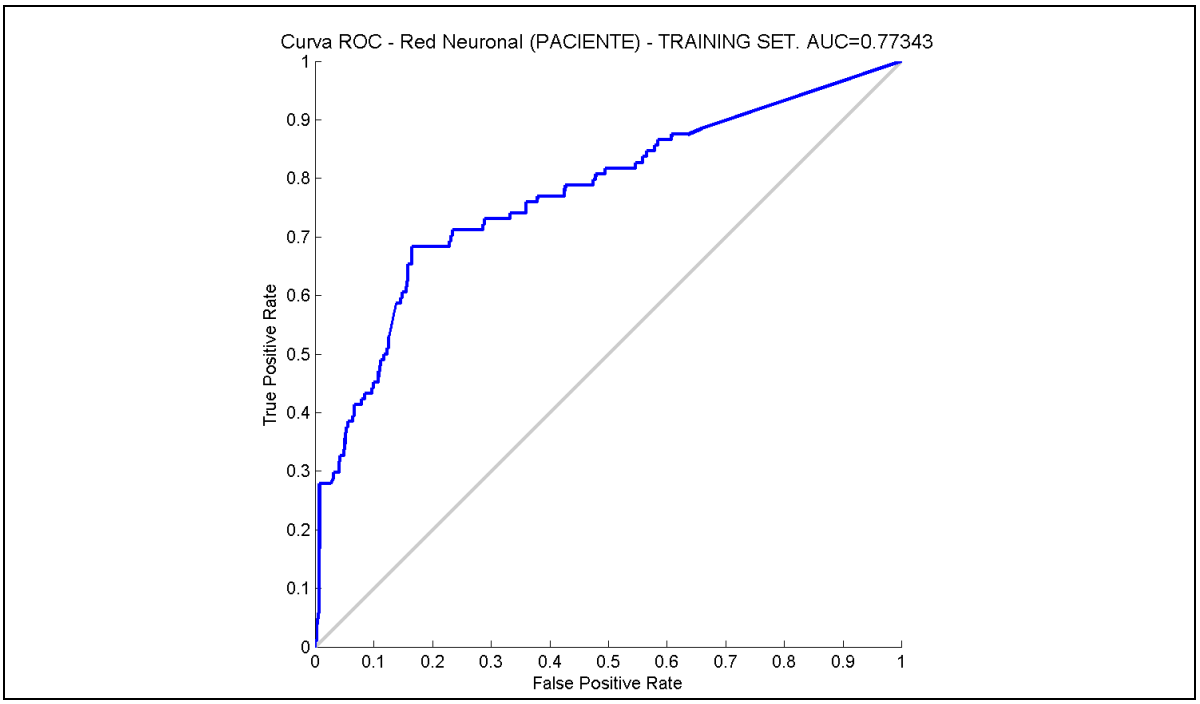
En la Tabla 21 se muestran todas las estadísticas de clasificación con dicha RNA para el modelo PACIENTE. Destaca la especificidad de 0.81, la sensibilidad de 0.60 y valor de G de 0.695 en el conjunto de evaluación. Puesto que la Regresión Logística y esta RNA fueron los clasificadores que mejor rendimiento tuvieron con esta versión reducida de los

datos, será necesario elegir cuál de los dos se ajusta más a nuestras necesidades para implementarlo en la aplicación móvil.

RED NEURONAL MODELO PACIENTE	Entrenamiento	Validación	Evaluación
G	0.755	0.753	0.695
AUC	0.773	0.749	0.755
SEN	0.683	0.708	0.594
ESP	0.836	0.801	0.814
VPP	0.350	0.315	0.292
VPN	0.953	0.955	0.939
ACC+IC	0.82 [0.80-0.84]	0.79[0.74-0.85]	0.79[0.74-0.84]
RV+	4.158	3.561	3.188
RV-	0.380	0.364	0.499
MC	$\begin{pmatrix} 71 & 132 \\ 33 & 672 \end{pmatrix}$	$\begin{pmatrix} 17 & 37 \\ 7 & 149 \end{pmatrix}$	$\begin{pmatrix} 19 & 46 \\ 13 & 201 \end{pmatrix}$

Tabla 21: Resultados de clasificación con la mejor red neuronal encontrada. Modelo PACIENTE.

En la Figura 32 se muestran las curvas ROC de la mejor Red Neuronal para el modelo PACIENTE, con los datos de entrenamiento, validación y evaluación. Todos los valores de AUC están entre 0.75 y 0.77, lo que indica una vez más que el modelo no está sobreajustado a los datos de entrenamiento, y que sus resultados de clasificación son moderadamente fiables.



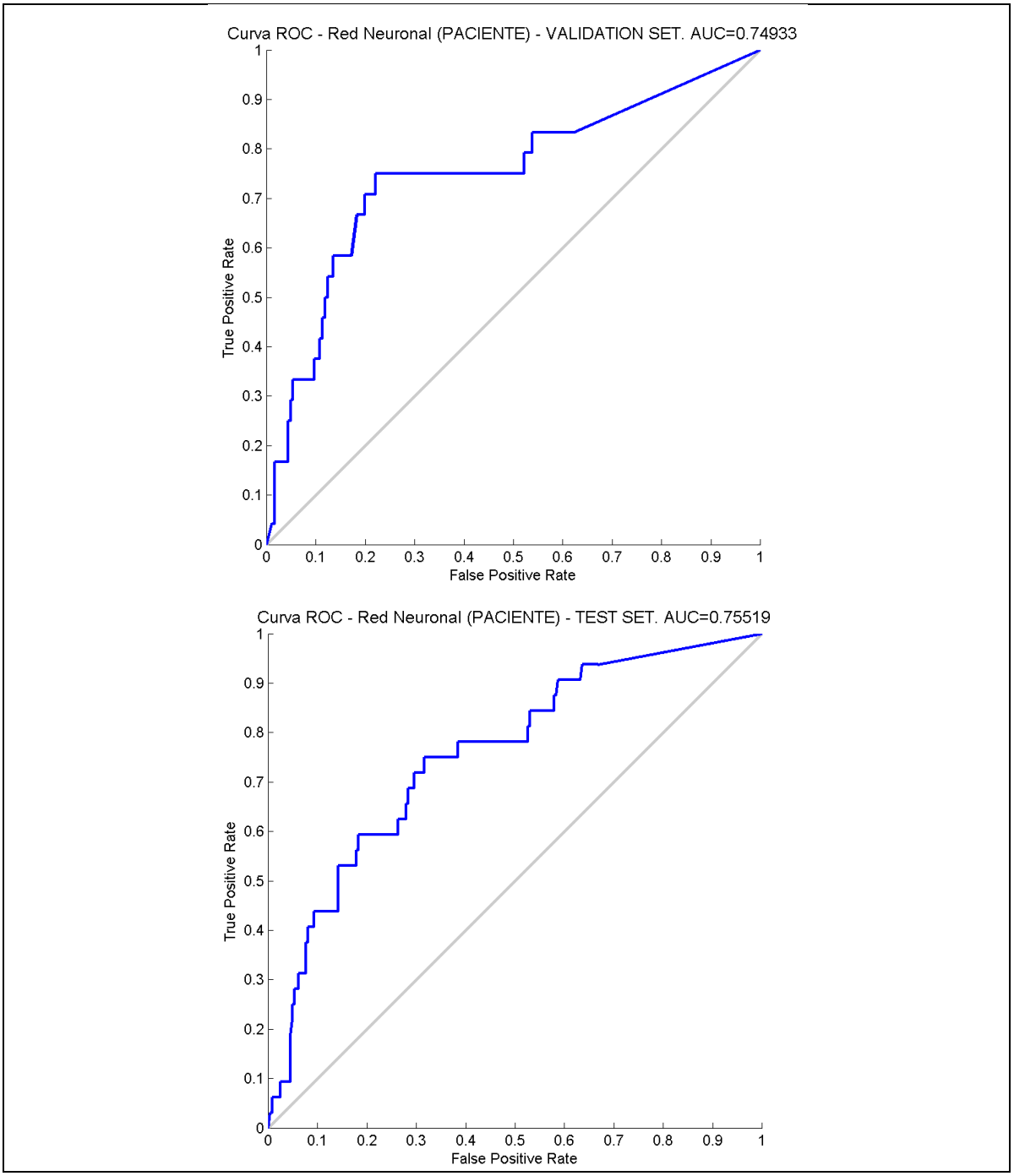
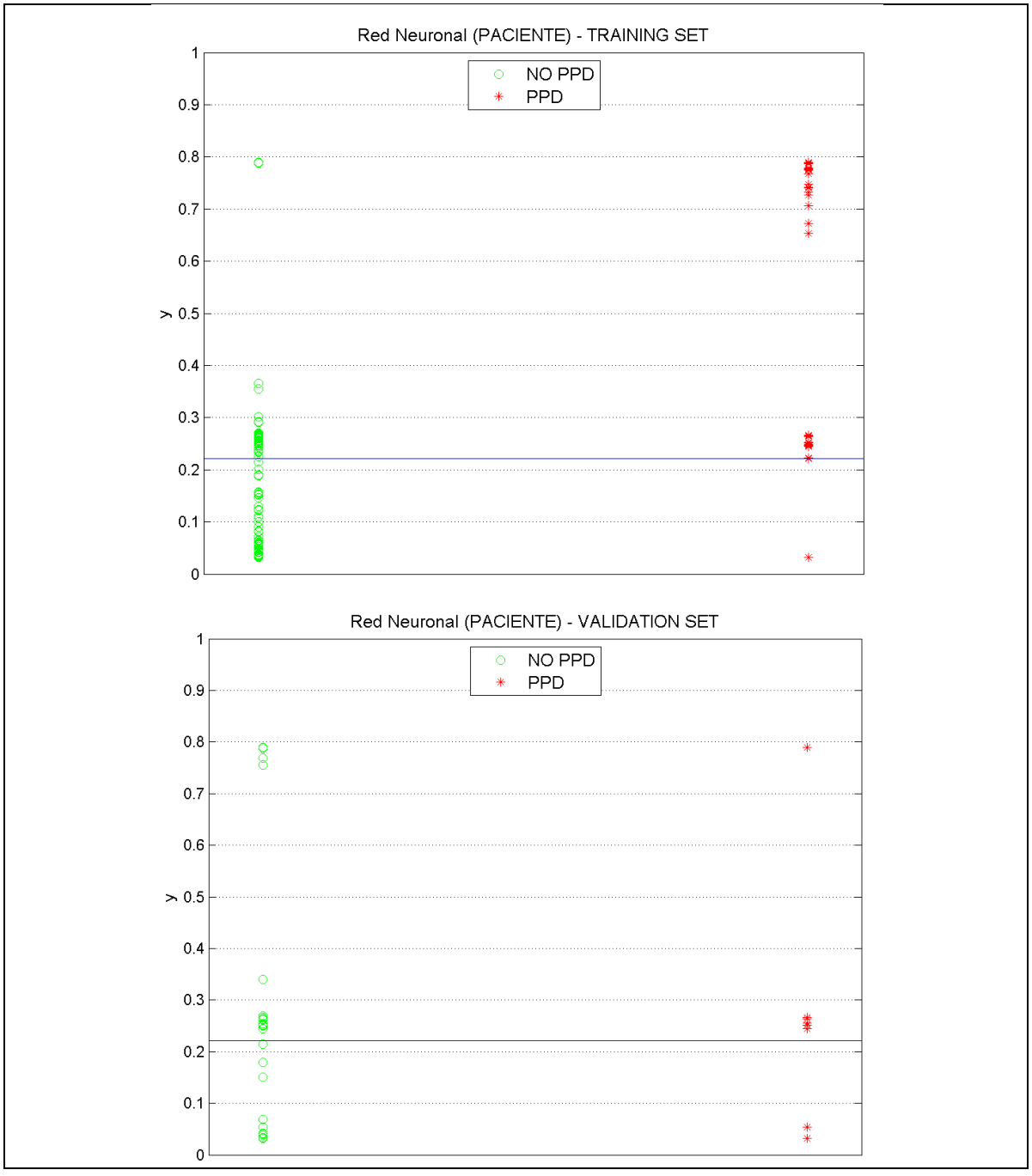


Figura 32: Curvas ROC obtenidas con la mejor red neuronal. Modelo PACIENTE.

La Figura 33 muestra la salida de la RNA y el umbral de decisión frente a la clase para el modelo PACIENTE, con los datos de entrenamiento, validación y evaluación. El umbral óptimo en este caso es de 0.22. En este modelo la población sana (NO DPP) está razonablemente bien ajustada en la parte baja del gráfico, es decir, en valores inferiores a 0.3, lo que propicia buenos resultados de especificidad (0.81).

Cabe destacar también que al contrario de lo que sucedía en la Regresión Logística, donde la población enferma (PPD) abarcaba prácticamente todo el espectro de la salida,

con esta RNA se concentra en unos pocos valores superiores al umbral. Esto es un punto a favor de dicha RNA a la hora de elegir el modelo que se implementará en la aplicación para dispositivos móviles. Sin embargo, aunque no se distribuye por todo el espectro de la salida, la población con DPP no termina de ocupar la parte alta del gráfico, es decir, en valores superiores a 0.8, lo que pudiera estar provocando la baja sensibilidad (0.60).



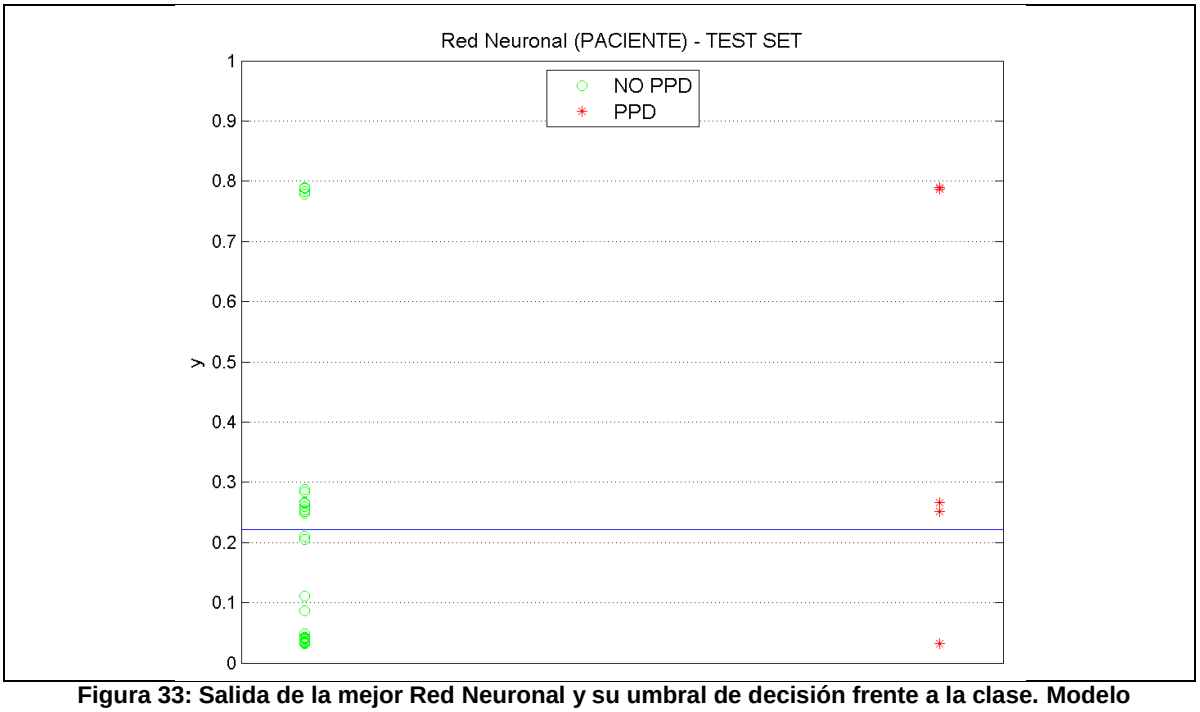


Figura 33: Salida de la mejor Red Neuronal y su umbral de decisión frente a la clase. Modelo PACIENTE.

4.4.1. Modelos Ensamblados

Es necesario mencionar que tanto para el modelo MÉDICO como para el PACIENTE, se realizó también un sistema de clasificación por votos de las mejores RNA obtenidas. Es decir, se escogieron las 3, 5 y 7 mejores RNA de cada experimentación, y se realizó el proceso de clasificación para cada conjunto de datos tal y como se ha descrito para la mejor RNA. Finalmente se escogía como resultado final la clase que mayor número de votos conseguía. Al ser un número de votos impar, una de las clases siempre era la ganadora. De esta manera se obtuvo una muy alta especificidad, pero sin embargo la sensibilidad caía demasiado y por lo tanto el valor de G nunca llegó a superar el de la mejor RNA, tanto con los datos referentes a la aplicación MÉDICA como a la PACIENTE. Puesto que una única RNA siempre mejoraba el sistema por votación, se optó por no incluir los resultados de dicha votación en este estudio.

4.4.2. Modelos Jerárquicos

También se llevó a cabo otra experimentación separando los datos de aquellas mujeres que superaban una puntuación de 9 en la variable EPDS de las que no. Para ambas particiones de datos se crearon todos los tipos de modelos descritos anteriormente (Naïve Bayes, Regresión Logística, SVM y RNA). La idea era mejorar la sensibilidad o especificidad de la mejor RNA obtenida al reforzar su resultado con algún modelo que tuviera un mejor rendimiento en uno de los dos grupos nuevos creados (EPDS > 9 o EPDS ≤ 9). Sin embargo, y tras todas las combinaciones de clasificadores y salidas

posibles, nunca se llegó a mejorar por encima de una centésima el valor de G de la mejor RNA en el modelo MÉDICO, y no se consiguió mejora alguna en el modelo PACIENTE.

Podemos concluir pues, que para este problema es mejor utilizar un único clasificador y no una variante de clasificación por votos o derivados, ya que no mejora sustancialmente los resultados de la primera opción.

4.5. Visión global de los resultados experimentales y modelos seleccionados para implementación en dispositivos móviles

Una vez presentados todos los tipos de clasificadores con los que se experimentó, en la siguiente sección se dará una breve visión global de sus resultados. Además, se seleccionarán aquellos con más rendimiento o que mejor se ajusten a nuestras necesidades para migrarlos a la plataforma móvil. Como criterio principal de evaluación se utiliza el valor de G sobre el conjunto de datos de evaluación, seguido por el AUC. En casos donde estos valores sean muy semejantes se podrán adoptar otros criterios, tal y como sucede en el modelo PACIENTE.

La Figura 34 muestra estos valores de G sobre el conjunto de datos de evaluación conseguidos con los clasificadores de tipo Naïve Bayes, Regresión Logística, SVM y RNA. Con un resultado de 0.80, la mejor RNA conseguida supera por 0.064 puntos de diferencia al segundo mejor, y por 0.13 al último. Además, también se ha de tener en cuenta que durante el estudio de Tortajada *et al.* [12], el mejor valor de G obtenido en sus modelos fue de 0.82. Teniendo en cuenta que en dicho estudio se contó con una variable independiente más, nuestro valor de G de 0.80 es un muy buen resultado. Se concluye por tanto que el clasificador a implementar en la aplicación móvil destinado a personal clínico será la mejor RNA.

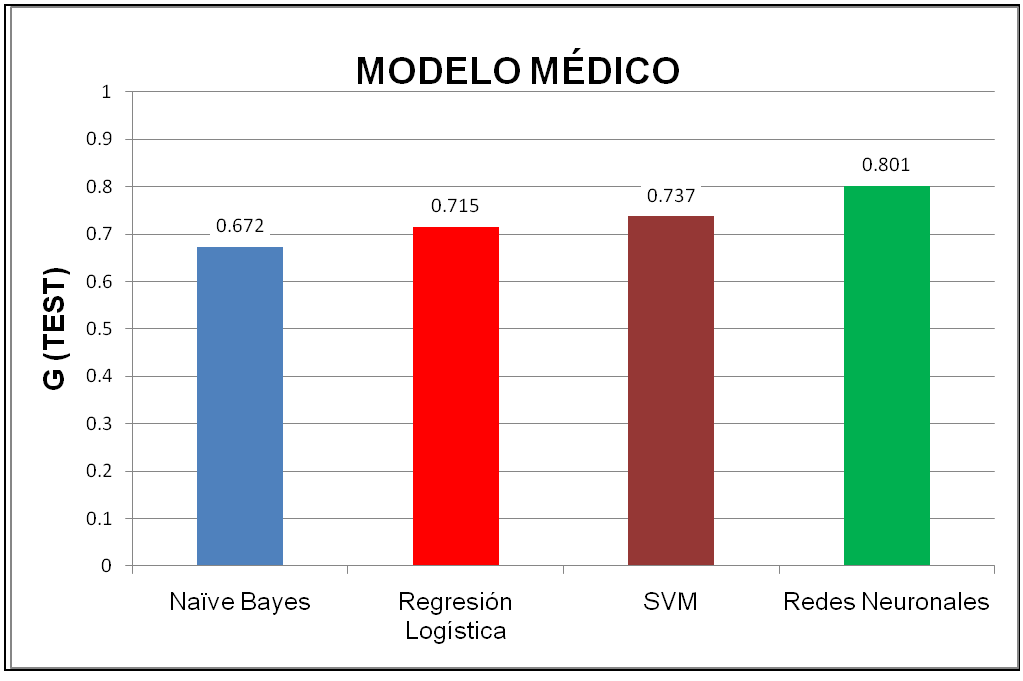


Figura 34: Resultados de G con los datos de test en los distintos modelos MÉDICO entrenados.

Respecto al modelo PACIENTE, la Figura 35 muestra estos mismos valores de G sobre el conjunto de evaluación. En este caso la selección del mejor clasificador es algo más complicada, ya que la Regresión Logística y la RNA presentan resultados muy parecidos

(0.706 y 0.695 respectivamente). La Regresión Logística presenta una AUC de 0.748 mientras que la RNA de 0.755. Ante estas condiciones de casi igualdad de prestaciones, hemos de tener en cuenta que este modelo estaría destinado a ponerse en manos de madres tras el parto mediante una aplicación móvil. En este caso es preferible no alarmar innecesariamente a las pacientes que la utilizasen, con lo que la prioridad es reducir todo lo posible en número de Falsos Positivos. La Regresión Logística, con 83 FP prácticamente dobla a la RNA con 46, siendo la especificidad de la primera del 0.66 y de la segunda del 0.81. Además, la población enferma no abarca todo el espectro de la salida en el caso de la RNA, tal y como sí sucede con la Regresión Logística. Teniendo todo esto en cuenta, finalmente se opta por implementar en la aplicación móvil destinada a pacientes el modelo conseguido mediante RNA.

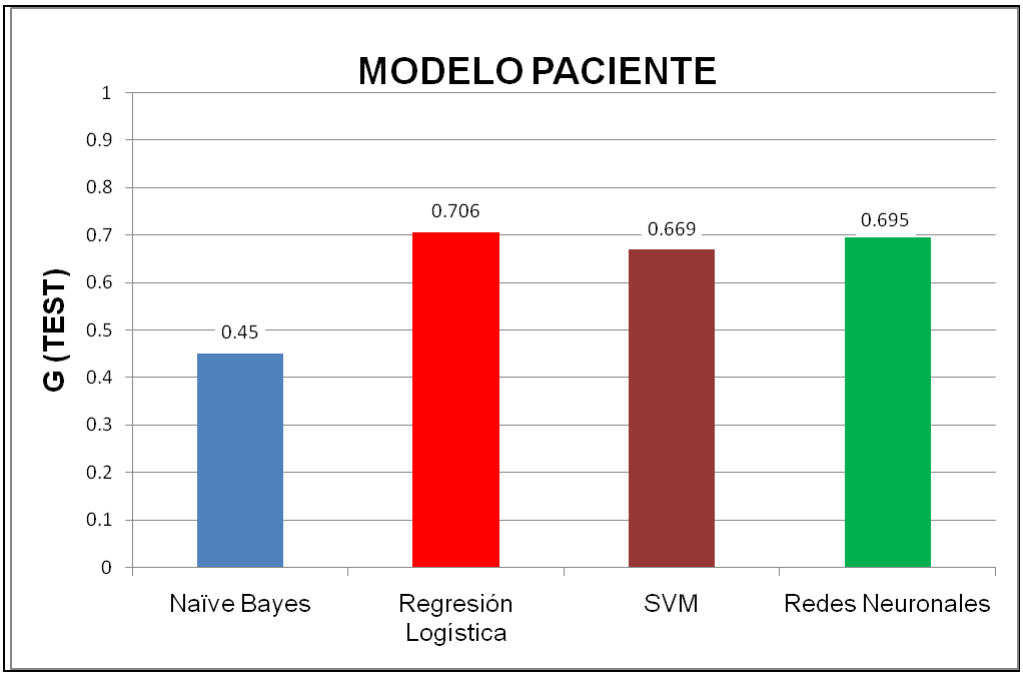


Figura 35: Resultados de G con los datos de test en los distintos modelos PACIENTE entrenados.

En la Tabla 22 se muestra una comparación de los resultados de nuestros mejores clasificadores en ambas versiones de la experimentación, frente a los resultados de la mejor RNA que consiguieron Tortajada *et al.* en su estudio con 16 variables independientes. Esta comparación se hace en base a los resultados obtenidos en el conjunto de evaluación. Ahí podemos ver como con el modelo MÉDICO, a pesar de contar con una variable independiente menos, hemos conseguido mejorar ligeramente el AUC, la especificidad y la tasa de aciertos. Sin embargo, los valores de G no mejoran debido a que nosotros tenemos una sensibilidad inferior en 0.06. En todo caso, podemos decir que se ha conseguido desarrollar un modelo con un rendimiento muy semejante, pero que necesita una variable menos de entrada.

Cabe destacar también que se consiguió una elevada especificidad de 0.81 en el modelo PACIENTE. De esta manera se espera un fallo mínimo al ofrecer un resultado negativo a

las madres sanas que utilicen nuestra aplicación y una tasa aceptable de fallo al dar resultados positivos.

COMPARACIÓN MEJORES MODELOS	Tortajada et al.	M. MÉDICO	M. PACIENTE
G	0.82	0.80	0.70
AUC	0.82	0.85	0.76
SEN	0.84	0.78	0.59
ESP	0.81	0.82	0.81
ACC+IC	0.81 [0.76-0.86]	0.82[0.77-0.86]	0.79[0.74-0.84]

Tabla 22: Comparación del mejor modelo obtenido por Tortajada et al. frente a los mejores conseguidos en este trabajo. Resultados sobre el conjunto de datos de evaluación.

Las comparaciones anteriores son en base a la misma base de datos, pero se desconoce si comparten idéntica división de las muestras en los respectivos conjuntos de entrenamiento, validación y evaluación. Esto es debido a que en ambos estudios la inclusión de cada muestra en un conjunto u otro se hizo de manera aleatoria, manteniendo las proporciones de las clases en todos.

Una vez ya seleccionados los clasificadores que se utilizarán en ambas aplicaciones móviles, el siguiente paso es reentrenarlos con todos los datos disponibles. Puesto que ambos son de tipo RNA, para este reentrenamiento se siguió el mismo proceso descrito anteriormente, pero esta vez utilizando como datos de entrenamiento toda la base de datos. Se utilizó las topologías y matriz de pesos iniciales W y bias b de las inicializaciones que finalmente dieron los mejores resultados. Se reentrenó otra vez en las mismas condiciones, es decir, mediante optimización *Levenberg-Marquardt backpropagation* por descenso de gradiente [66], con condiciones de parada del aprendizaje de un máximo de 100 *epochs* y un incremento mínimo del gradiente de 0.0001 entre dos *epochs* consecutivos. De este modo se reestimaron dichas matrices de pesos W y bias b en las RNA para los modelos MÉDICO y PACIENTE.

Al igual que anteriormente, se seleccionó el mejor umbral o punto de corte para clasificar entre madres con y sin riesgo de DPP seleccionando aquel que maximizase la función G, pero esta vez con todos los datos disponibles. En el modelo MÉDICO el umbral resultó ser de 0.49 y de 0.15 en el PACIENTE.

Se considera que el rendimiento alcanzado con este reentrenamiento es elevado y coherente con los resultados obtenidos durante la fase de experimentación. Puesto que existen más datos en esta ocasión durante el entrenamiento, los modelos ajustan mejor las separaciones entre clases. El no obtener un resultado perfecto nos indica que no existe un sobreajuste sobre los datos de entrenamiento y que por lo tanto la generalización debería ser correcta.

La Figura 36 muestra la salida, y el histograma de la salida, de la RNA del modelo MÉDICO con todas las muestras de la base de datos respecto a sus respectivas clases. También se representa el umbral de decisión o punto de corte entre ambas clases mediante una línea perpendicular al eje de la salida. Como ya se ha comentado, el umbral óptimo en el modelo MÉDICO es de 0.49, valor que además de dejar por debajo a

casi todas las madres sin DPP, indica la buena calibración del modelo. También se aprecia como las madres con DPP aparecen mayoritariamente en valores superiores a 0.80. En el histograma vemos el número de muestras que se acumulan en cada región de la salida, siendo los aciertos los que quedan en la parte izquierda del umbral en la clase NO DPP, y los que quedan a la derecha en la clase DPP. De este modo podemos ver la importancia de cada tramo de la salida. Se aprecia que existen 43 falsos negativos que se acumulan en la misma zona que los verdaderos negativos. Seguramente esto es debido a la semejanza entre esas madres que finalmente desarrollaron DPP y las que no, por lo que al algoritmo de aprendizaje se le hace muy difícil distinguirlas. Sin embargo vemos como este clasificador ajusta muy bien la clase NO DPP en valores muy cercanos al 0, no existiendo prácticamente ningún falso positivo. Las madres con DPP en la parte de verdaderos positivos se acumulan mayoritariamente en valores cercanos al 1.

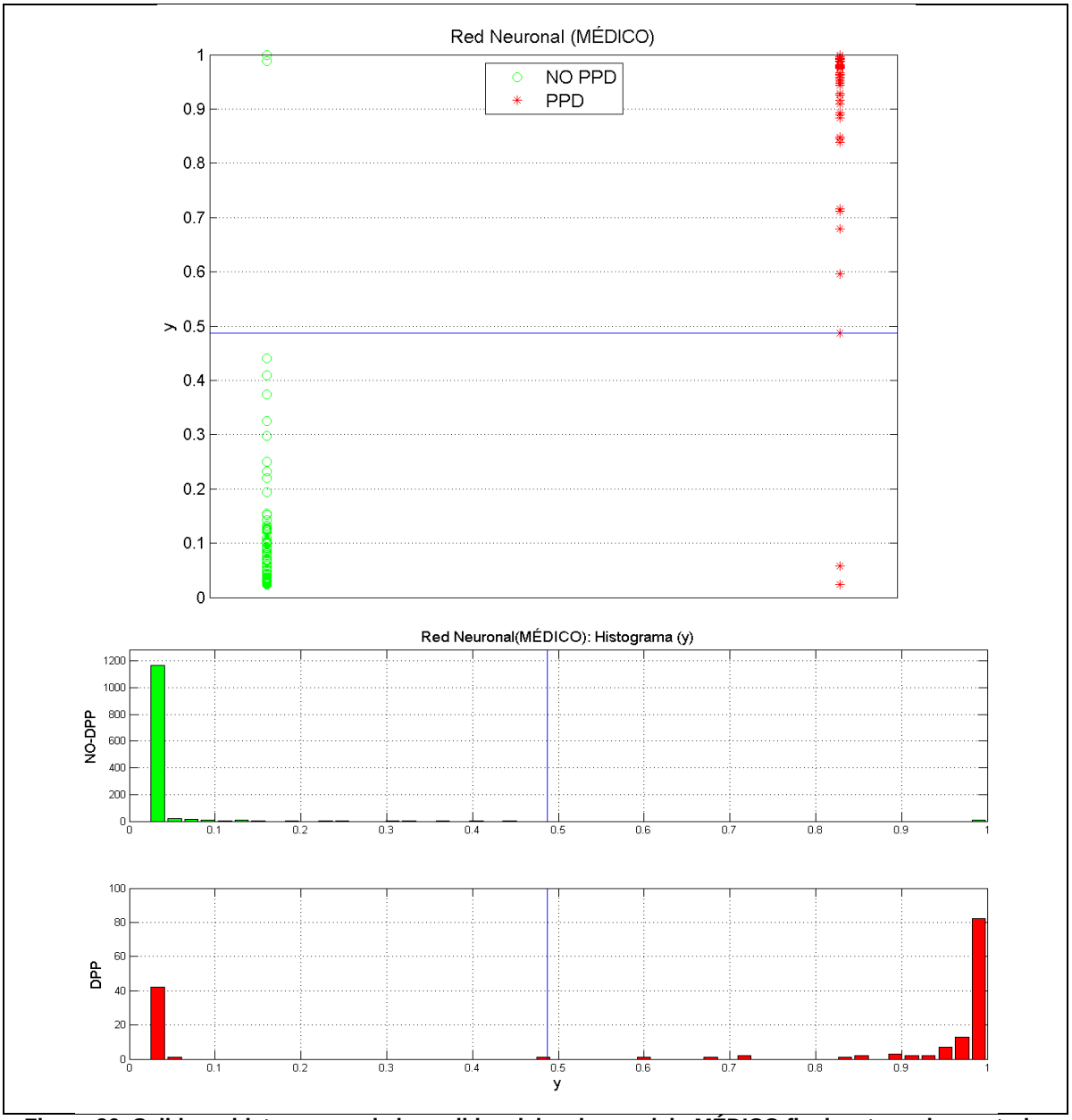


Figura 36: Salidas e histogramas de las salidas del mejor modelo MÉDICO final, entrenado con todos los datos, junto su umbral de decisión, frente a la clase.

En la Figura 37 muestra la salida, y el histograma de la salida, de la RNA del modelo PACIENTE con todas las muestras de la base de datos respecto a sus respectivas clases. El umbral de decisión se representa igual que en las figuras anteriores, tomando en este caso un valor de 0.15. Se aprecia que prácticamente toda la población sana queda por debajo del 0.5 en la salida del clasificador, pero sin embargo la población con DPP sí que aparece en casi todo el espectro de la salida. Esto provoca que el umbral se sitúe en 0.15. No obstante, el resultado general conseguido con este modelo con una especificidad de 0.80 cumple nuestras necesidades a la hora de desarrollar la aplicación móvil para madres, tal y como se explicó anteriormente, manteniendo una sensibilidad de 0.70.

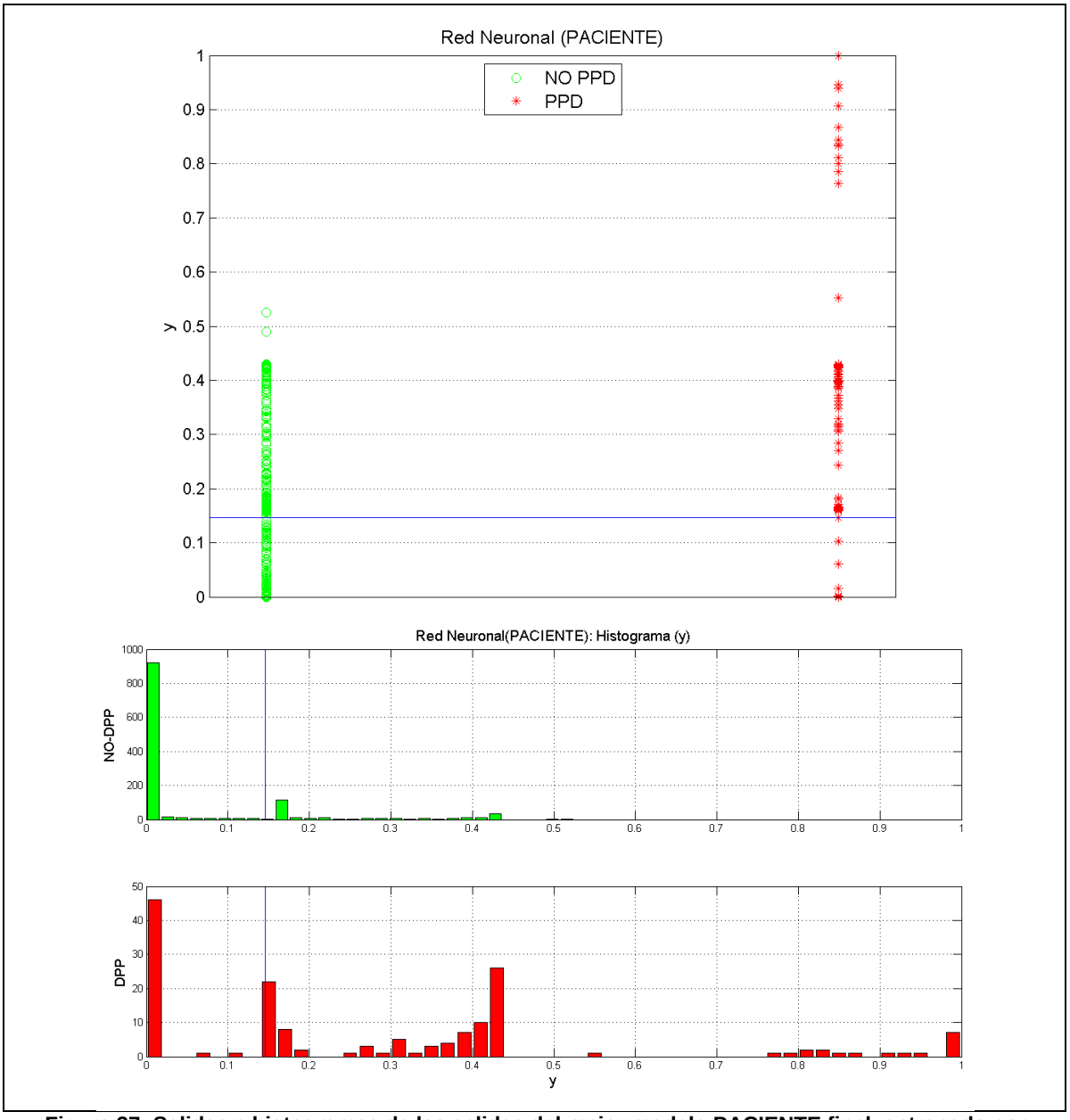


Figura 37: Salidas e histogramas de las salidas del mejor modelo PACIENTE final, entrenado con todos los datos, junto su umbral de decisión, frente a la clase.

Se tendrán en cuenta estas últimas figuras, puesto que se necesita que nuestra aplicación dé no sólo un resultado binario de posible riesgo de desarrollo de DPP ante una nueva clasificación, sino que se requiere también ofrecer al usuario una medida de probabilidad o certeza. Para ello se desarrollaron las ecuaciones de la Figura 38 donde θ es el umbral de la RNA para clasificar entre DPP y NO DPP, mientras que $\hat{y}$ es la salida de la RNA ante la clasificación de una nueva muestra. Por otro lado $N(y_{NO\_DPP} \geq \hat{y})$ es el número de participantes en el estudio sin DPP cuya salida y en la red neuronal supera el valor de $\hat{y}$. N_{NO_DPP} es una constante determinada por el número total de madres del estudio que finalmente no desarrollaron DPP, es decir, 1237. Es fácil ver cómo el posible error que estaremos cometiendo al clasificar de este modo a una nueva madre en el grupo de DPP puede expresarse mediante un valor entre 0 y 1 expresado como $error(c_{DPP}|\hat{y} \geq \theta)$. Al dividir el número de madres sanas del estudio que la RNA clasificó mal y que además superan el valor de $\hat{y}$ entre el número total de pacientes sin DPP (1237) obtendremos el posible error de clasificación esperado con nuestra base de datos en la clase DPP.

Lo mismo sucede, pero a la inversa, si queremos conseguir el error cometido en la clasificación de una nueva muestra en el grupo de las mujeres sin DPP: $error(c_{NO\_DPP}|\hat{y} < \theta)$. En este caso tendremos que dividir el número de mujeres enfermas que no superaron el valor de $\hat{y}$, denotado por $N(y_{DPP} < \hat{y})$, entre el número total de pacientes con DPP, denotado por N_{DPP}. Hemos de recordar que este número total de pacientes enfermas en nuestro estudio fue de 160.

$$error(c_{DPP}|\hat{y} \geq \theta) = \frac{N(y_{NO\_DPP} \geq \hat{y})}{N_{NO\_DPP}} \qquad error(c_{NO\_DPP}|\hat{y} < \theta) = \frac{N(y_{DPP} < \hat{y})}{N_{DPP}}$$

Figura 38: Medidas de error al clasificar una nueva muestra que utilizaremos en la aplicación móvil.

Tanto en la red neuronal final relativa a la aplicación para personal clínico, como en la relativa a la aplicación para pacientes, se hizo un barrido de todos los posibles valores de la salida $\hat{y}$ para calcular esta función de error, tal y como se aprecia en la Figura 39. Para ambos clasificadores se representan sus respectivos umbrales de decisión y función de error. Hemos de tener en cuenta que la parte a la izquierda del umbral muestra la zona de clasificación de madres sanas o NO DPP, mientras que la de la derecha indica clasificación como enfermas o DPP.

La interpretación de estos gráficos es que en ambos modelos siempre existe cierta incertidumbre a la hora de dar un resultado sobre pacientes sanas, pero sólo el modelo PACIENTE presenta también esa incertidumbre a la hora de clasificar enfermas en el rango de su salida desde el umbral 0.15 hasta valores de 0.5. A partir de una salida $\hat{y}$ de 0.5 los dos clasificadores prácticamente aseguran que la paciente es población de riesgo para padecer una DPP.

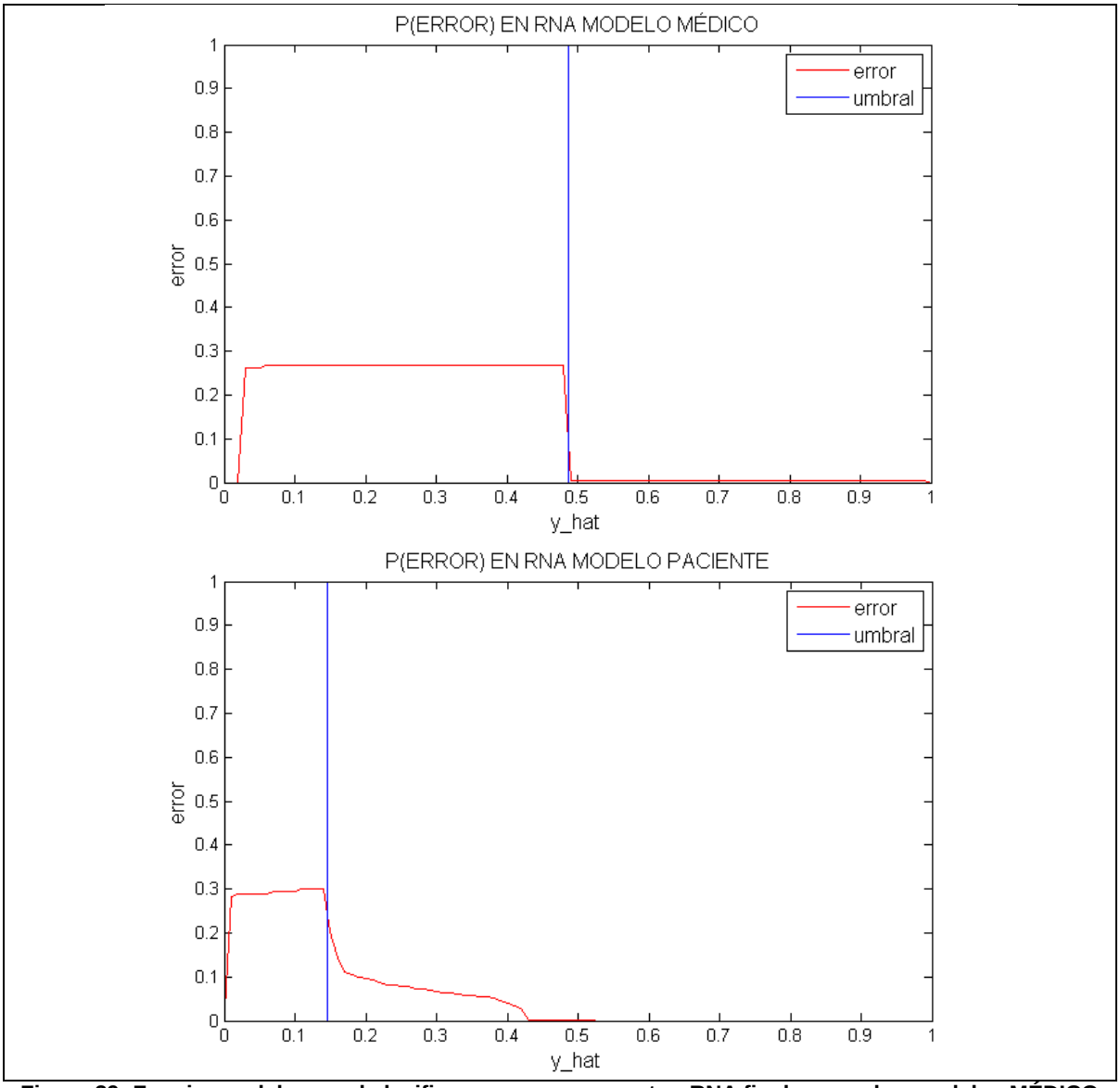


Figura 39: Funciones del error al clasificar una nueva muestra. RNA finales para los modelos MÉDICO y PACIENTE.

Una vez descritos los modelos finales seleccionados, sus resultados y funciones de error, en la siguiente sección se explicarán todos los detalles relacionados con la transferencia de estos clasificadores a las aplicaciones móviles correspondientes.

5. Aplicación para dispositivos móviles desarrollada

Una vez seleccionadas las redes neuronales para el modelo MÉDICO y para el modelo PACIENTE, con sus parámetros ya estimados, se mostrará durante esta sección la aplicación para dispositivos móviles creada que contempla ambos casos. En el primero el usuario sería personal clínico capaz de conocer y saber interpretar todas las variables del modelo MÉDICO (incluidas N-exp-8s, N-exp-32s, EPQN, DUKE-suma y 5-HTT-GC), pudiendo introducirlas como entrada en la aplicación para obtener el resultado de clasificación de una paciente. En el segundo caso, el usuario sería una madre que se encuentra en su primera semana tras dar a luz y quiere conocer si es población de riesgo para desarrollar una DPP. En este último caso, evidentemente las variables clínicas mencionadas anteriormente no podrían ser introducidas por ella ya que las desconocería, y por lo tanto sería necesario utilizar el modelo que durante todo este tiempo hemos llamado PACIENTE, donde se excluyen.

Es por lo tanto necesario, dadas estas condiciones iniciales que se plantean, establecer unos requisitos mínimos para la aplicación.

El primero de estos requisitos es que la aplicación funcione en dispositivos móviles, tales como *smartphones* o tabletas. Esta idea viene como consecuencia de que estos nuevos terminales ofrecen unas capacidades similares a las de un ordenador personal, pero a diferencia de ellos, siempre están en los bolsillos de sus usuarios. Esto permite un abanico de aplicaciones mucho más cercanas al usuario. En nuestro caso, una madre que acabe de dar a luz y esté convaleciente durante los primeros días en el hospital, seguramente tendrá su teléfono móvil cerca. En el momento oportuno, podría contestar las preguntas que le planteamos en la aplicación en pocos minutos y de esta manera conocer si es población de riesgo para padecer DPP. Por lo tanto, una aplicación de este tipo podría servir como una herramienta muy potente en la prevención de la enfermedad. Lo mismo sucede con el personal clínico, el cual se beneficiaría de poder comprobar en un dispositivo más cómodo y cercano, el riesgo de DPP que tienen sus pacientes. De este modo estaríamos ayudando en su proceso de toma de decisiones al diagnosticar un nuevo caso de la enfermedad.

El segundo requisito que se establece es que la aplicación, como mínimo, debe funcionar en una plataforma de amplia difusión, para de este modo poder llegar al mayor número posible de usuarios potenciales. Tras evaluar todas las opciones, se decidió utilizar la plataforma Android ya que, además de su amplia implantación en el mercado, provee potentes herramientas de desarrollo de libre acceso. Otra ventaja que presenta Android es que si queremos publicar una nueva aplicación en su tienda más conocida, *Google Play Store*, no requerimos de ninguna revisión por parte de nadie para poder ofrecerla al público en general. No ocurre lo mismo con las otras dos plataformas más extendidas. *Apple* controla el contenido de lo que se publica en su *Apple Store* en lo que respecta a sus dispositivos *iPhone* y *iPad*, pudiendo rechazar una nueva aplicación si lo desea. *Microsoft* hace lo mismo en su *Microsoft Store* de aplicaciones para *Windows Mobile* [68]. Esto no quiere decir que en un futuro no sea recomendable hacer funcionar nuestra aplicación móvil en estas dos últimas plataformas, pero para este prototipo inicial, Android

nos ofrece una mayor flexibilidad y mayor número de acceso a posibles usuarios. Según cifras oficiales, actualmente se activan más de un millón de dispositivos con Android diariamente [69]. La versión de este sistema operativo sobre la que funciona la aplicación desarrollada es Android 2.1, la cual apareció en enero de 2010. Hay que tener en cuenta que cualquier versión posterior a la 2.1 también soportará nuestra aplicación.

El siguiente requisito es de tipo funcional. Consiste en que nuestra aplicación presente claramente cuál es su objetivo, permita responder al usuario las preguntas que le planteemos de forma sencilla y sin ambigüedades, y le presente los resultados de clasificación de manera entendible. Esto tiene que ver con el concepto de usabilidad del *software*, donde una interfaz agradable consigue captar mejor la atención de quien utiliza la aplicación. Por el contrario, otro software que realice las mismas tareas pero cuya interfaz no sea entendible por sus usuarios finales suele generar rechazo, produciendo el abandono de su uso o generando resultados incorrectos debido a imputaciones de datos erróneas. Por lo tanto, el requisito de una buena usabilidad ha de ser tenido en cuenta.

Una vez descritos los requisitos de nuestra aplicación, es necesario hacer una breve descripción de las estructuras de datos creadas para manejar la información que utilizaremos. El lenguaje de programación que se utiliza para generar aplicaciones Android es Java, el cual está basado en el paradigma de la programación orientada a objetos. Un objeto no es más que una estructura de datos que puede contener los valores de distintas variables de tipo numérico o de texto entre otros, llamados atributos. Sobre estos atributos se pueden aplicar todo tipo de operaciones, a las que se les conoce por funciones o métodos.

En primer lugar, hay que tener en cuenta que, ya sea a personal clínico o madres, nuestra aplicación les realizará un cuestionario. Un *Cuestionario* será un objeto compuesto por una lista de preguntas.

A su vez, una *Pregunta* será un objeto compuesto por el texto destinado a personal clínico, el texto destinado a pacientes, y el valor correspondiente a la respuesta que haya dado el usuario. Además, si una pregunta es de tipo categórico, mantendrá una lista con los textos que corresponden a cada una de las opciones a escoger. Una pregunta puede no aplicar al modelo PACIENTE, por lo que en tal caso no ha de aparecer en ese *Cuestionario*.

Una vez estén todos los valores de las respuestas de un cuestionario introducidos, cada uno de esos valores ha de ser codificado siguiendo las mismas normas que durante el entrenamiento de los modelos de clasificación, conforme a la variable correspondiente a cada pregunta. A los objetos encargados de codificar estos valores y generar el vector de entrada para el *Clasificador* correspondiente les llamaremos *Codificador*.

Finalmente, los últimos y más importantes objetos, serán los llamados *Clasificador*, los cuales, ante un vector con las entradas codificadas, ofrecerán el resultado de su clasificación como un vector donde se indique el valor de salida estimado $\hat{y}$, la clase correspondiente a ese valor, y el error estimado ante tal predicción.

Es fácil comprender que fueron estos últimos objetos de tipo *Clasificador*, sobre los que se codificaron las mejores redes neuronales resultantes de la experimentación, tanto la

correspondiente al modelo MÉDICO, como al modelo PACIENTE. Es decir, se migró de la plataforma de experimentación (Matlab 2013a) al lenguaje de programación Java la matriz de pesos W y vector de *bias* b correspondientes a cada modelo, además de codificar el comportamiento de cada red neuronal con su respectiva topología.

Una fase muy importante fue la validación exhaustiva con toda la base de datos de que, ante un mismo vector de entrada, se obtenía el mismo resultado tanto en la plataforma de experimentación, como con las estructuras de datos Java que se implementaron. De esta manera, nos aseguramos unívocamente que los modelos eran los mismos y presentaban el mismo comportamiento que durante la fase de experimentación.

Una de las ventajas de implementar nosotros mismos las redes neuronales en el código Java, incluyendo las matrices de pesos y *bias*, sin pasar por librerías de terceros, es que han presentado una velocidad realmente elevada en su utilización en dispositivos móviles. En un principio se tuvo miedo a que la cantidad de cálculos a hacer fuera a ser un problema a la hora de usar estos terminales. Sin embargo, fue una agradable sorpresa comprobar que el cálculo de las combinaciones lineales y exponenciales necesarias para la función tansig, no fue apreciable en tiempo ni en el emulador de Android, ni en *smartphones* relativamente antiguos como el HTC Desire HD.

Una vez presentados los requisitos, estructuras de datos básicas creadas, y correcta validación del funcionamiento de las RNA implementadas en Java, nos queda presentar la Interfaz Gráfica de la aplicación móvil desarrollada en el siguiente apartado.

5.1. Interfaz Gráfica de la aplicación móvil desarrollada

En este apartado se dará una visión de la Interfaz Gráfica de Usuario (IGU) de la aplicación móvil Android desarrollada, además de explicar cómo un usuario podrá interactuar con ella.

Para empezar, hemos de mencionar que todas las aplicaciones en Android y cualquier otro sistema operativo tienen un nombre que las identifica. A partir de este momento utilizaremos *'eDPP Predictor'* para referirnos a nuestra aplicación, ya que es el nombre que se eligió en nuestro caso, siendo la '*e*' acrónimo de *electronic*.

El principal cometido de la IGU de una aplicación consiste en proporcionar un entorno visual sencillo para permitir la comunicación entre dicho software y el usuario que lo utiliza. En Android, el concepto de *Activity* representa una unidad de interacción con el usuario, correspondiendo a lo que coloquialmente llamamos una pantalla de la aplicación. Una aplicación suele estar formada por una serie de actividades, de forma que el usuario puede ir navegando entre ellas. Además, en Android se dispone de un botón (físico o en pantalla) que permite volver a la *Activity* anterior. Todas las actividades que conforman una aplicación deben perseguir un objetivo común. En nuestro caso, ese objetivo es que el usuario, ya sea personal clínico o una madre que ha dado a luz recientemente, conteste nuestro cuestionario y vea de una forma clara y entendible el resultado del clasificador, es decir, si la mujer es o no población de riesgo para sufrir una DPP.

En la Figura 40 se muestra el diagrama de flujo entre las distintas actividades que componen nuestra aplicación. Mediante este esquema se puede ver rápidamente la cantidad de actividades que tendremos además de la interacción del usuario final con la aplicación.

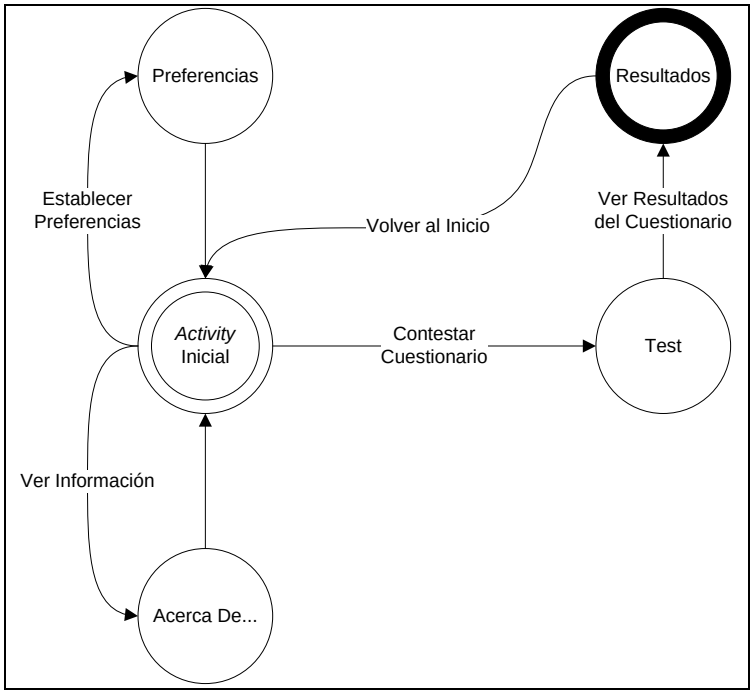


Figura 40: Diagrama de flujo entre las distintas *Activities* de *'eDPP Predictor'*.

La *Activity* Inicial es el punto de unión entre todas las demás, siendo la primera que aparecerá en pantalla a modo de menú principal. Desde ella podremos acceder a cambiar las preferencias, contestar el cuestionario o ver el clásico formulario 'Acerca De' con más información de la aplicación.

Respecto a la *Activity* de Preferencias de Usuario, es aquella donde podremos indicar el modo en que queremos que trabaje la aplicación, pudiendo ser '*Modo Clinic*' o '*Modo Madres*'. El primero hará que se trate al usuario con un lenguaje más técnico durante el cuestionario, además de pedir en este los valores de todas las variables independientes (incluidas N-exp-8s, N-exp-32s, EPQN, DUKE-suma y 5-HTT-GC) para poder utilizar el modelo MÉDICO como clasificador. El segundo modo tratará al usuario con un lenguaje más coloquial, suponiendo que es una madre en su primera semana tras dar a luz, por lo que no preguntará por las cinco variables clínicas anteriores. En el *'Modo Madre'* se utilizará el modelo PACIENTE como método para determinar si quien responde al test es o no población de riesgo de padecer DPP.

La diferencia más importante entre estos dos modos será que en *Clinic* pedirá el valor del test EPDS como si de una variable numérica se tratara, mientras que en *'Modo Madre'* se realizarán las 10 preguntas de dicho test y se calculará la puntuación EPDS en base a esas respuestas. Las preferencias del usuario quedarán guardadas en la memoria del dispositivo móvil, recordándose en posteriores usos de la aplicación.

La Figura 41 muestra las capturas de pantalla correspondientes a la *Activity* Inicial, 'Preferencias de Usuario' y 'Acerca De'. Tal como se apreciará en las sucesivas capturas, en caso de estar en *'Modo Clinic'* el título que aparecerá en la pantalla comenzará por *'eDPP Predictor - Clinic'*, y en el caso contrario por *'eDPP Predictor - Madres'*.

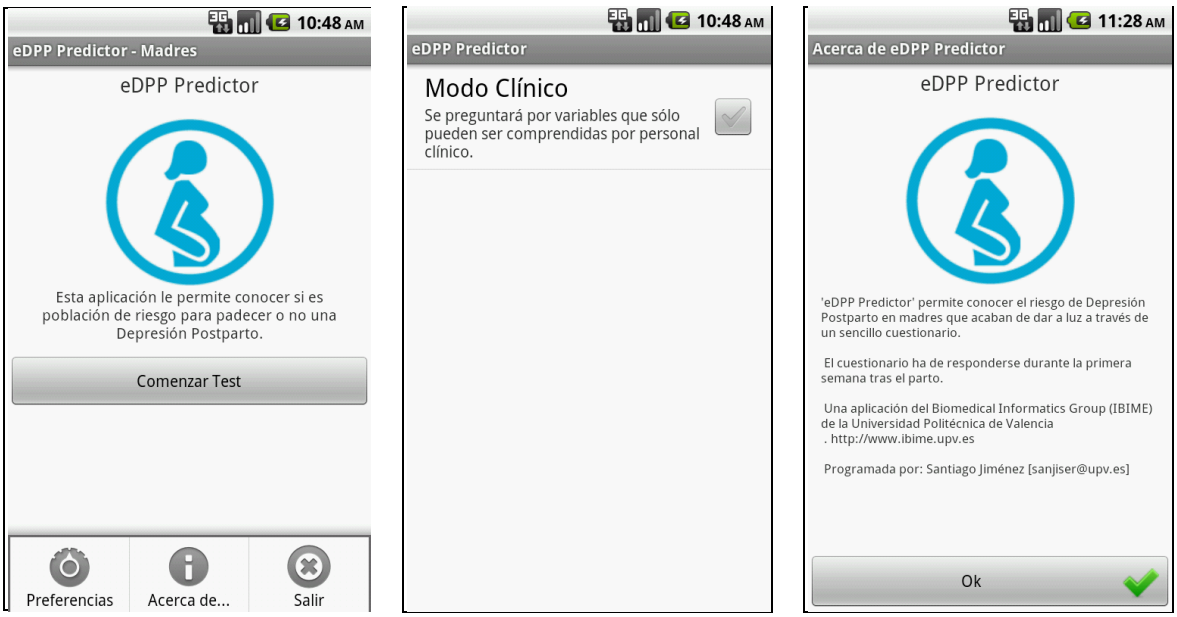


Figura 41: Capturas de pantalla de '*eDPP Predictor*': *Activity* Inicial, 'Preferencias de Usuario' y 'Acerca De'.

En la *Activity* Test se mostrará una a una las preguntas correspondientes a cada variable independiente del modelo MÉDICO o PACIENTE, dependiendo del modo en que nos encontremos. Puesto que las variables independientes que manejamos son de tipo categórico o numérico, tendremos dos posibles tipos de entrada para cada respuesta de nuestro test. La Figura 42 muestra diferentes capturas de pantalla de '*eDPP Predictor*' durante el cuestionario en modo 'Madres' y 'Clinic', donde se ilustran las dos posibles formas de entrada de datos por parte del usuario. En el caso de las variables categóricas, debajo de la pregunta o frase descriptiva correspondiente, aparece una lista de opciones sobre las que se debe seleccionar una. Esta selección se hace con una simple pulsación sobre la opción elegida para seguidamente pasar a la siguiente pregunta. En el caso de las variables numéricas, aparece una caja de texto sobre la que se introduce su valor mediante el teclado virtual.

Nótese que en el 'Modo Clinic' se permite avanzar a la siguiente pregunta marcando algunos valores como desconocidos, mientras que en el 'Modo Madres' no. De esta manera se consigue que las madres respondan a todas las cuestiones, mientras que el personal clínico puede dejar vacíos algunos de estos valores. El codificador de los valores de entrada al clasificador tratará estos valores desconocidos tal y como se hizo durante la fase de preprocesado de datos. Esta imputación de valores perdidos consistía en no activar ninguna de las unidades correspondientes a una variable categórica, poner el valor de la moda en variables discretas y el de la media en variables continuas. Los valores de moda y media en este caso serán los relativos a los datos con los que se entrenó el clasificador. Los valores desconocidos en las respuestas suponen una potencial pérdida en la precisión de la clasificación, pero en la enorme casuística de los usuarios clínicos serán inevitables en muchas ocasiones.

Recordar que en 'Modo Madres', la puntuación de la variable EPDS se conseguirá realizando las 10 preguntas de este test. Cada una de ellas es una cuestión sobre el estado de ánimo de la madre, donde se ha de seleccionar una opción de entre 4 posibles.

Figura 42: Capturas de pantalla de *'eDPP Predictor'*: Preguntas del cuestionario respecto a variables categóricas y numéricas. Versión 'Madres' y 'Clinic'.

Tras contestar a todas las preguntas del cuestionario, ya sea en modo 'Madres' o 'Clinic', la aplicación codifica todas las respuestas para la entrada del modelo correspondiente, es decir, la RNA del modelo PACIENTE en el primer caso, o la RNA del modelo MÉDICO en el segundo. Se obtiene el resultado de la clasificación de esta nueva muestra como perteneciente o no a población de riesgo de padecer DPP tras el parto.

Además también se obtiene la probabilidad de acierto, que no es más que $1 - error$, siendo $error$ la función descrita en el apartado final de la experimentación con clasificadores. En dicha experimentación, para todo el rango de posibles salidas del clasificador $\hat{y}$, donde $\hat{y} \in [0,1]$, se calculó el valor de la función $f(\hat{y}) = error$ para cada punto en intervalos de 0.01. Esta matriz $[\hat{y}, error]$ se incluyó directamente en el código fuente correspondiente a cada RNA. De este modo, para un valor concreto de $\hat{y}$ en un clasificador concreto, se puede obtener el valor de su $error$ precalculado. Cabe recordar que cada clasificador tiene una función de $error$ diferente, presentando el modelo MÉDICO un mejor rendimiento que el PACIENTE, por lo que podemos esperar más fiabilidad de la aplicación en caso de ser usada correctamente en modo '*Clinic'*.

Finalmente se ha de presentar este resultado de clasificación al usuario de una forma clara y entendible. Tal y como se ve en la Figura 43, la *Activity* de Resultados mostrará la predicción mediante dos cajas de texto. La primera presentará un mensaje donde se podrá leer 'POBLACIÓN DE RIESGO: SÍ', o 'POBLACIÓN DE RIESGO: NO' según el resultado positivo o negativo que haya dado el modelo. La segunda caja de texto mostrará la probabilidad de acierto descrita. De esta manera se da la información justa y necesaria para entender el resultado de nuestros métodos, evitando confundir al usuario con términos que quizás no llegase a entender.

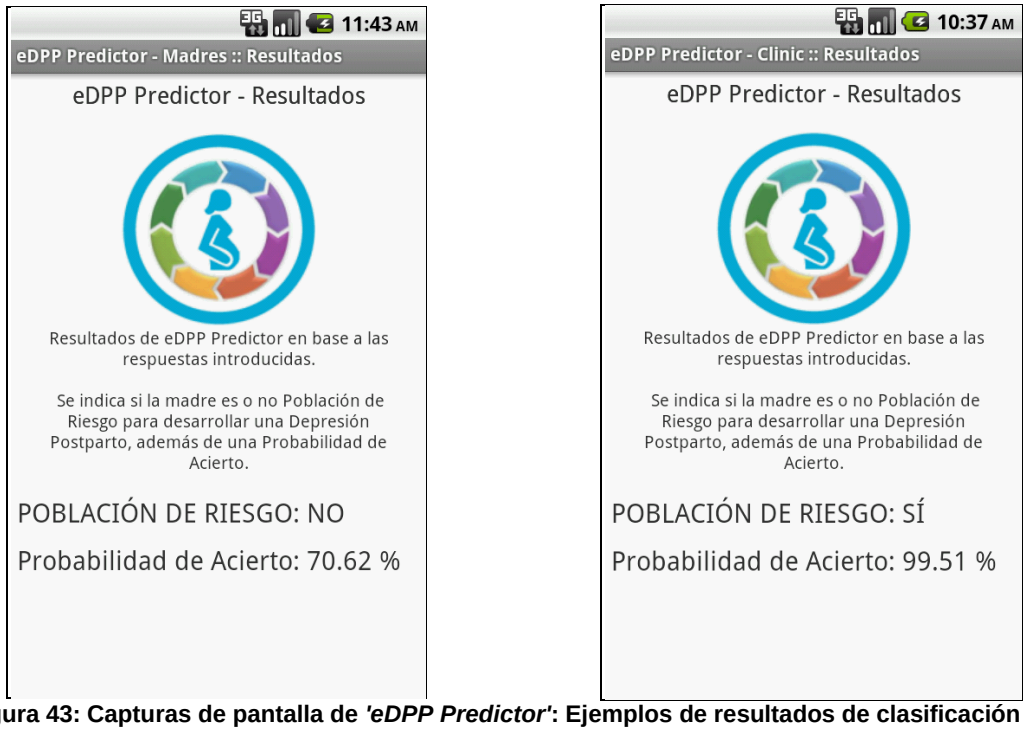


Figura 43: Capturas de pantalla de *'eDPP Predictor'*: Ejemplos de resultados de clasificación tras responder al test. Versión 'Madres' y 'Clinic'.

6. Conclusiones

En la presente sección se analizan las principales conclusiones que se extraen tras evaluar los experimentos y resultados obtenidos en los apartados anteriores, realizándose también un análisis crítico de sus limitaciones.

Tras la revisión del estado del arte en técnicas y métodos para la detección temprana de la DPP, se encontró que, a excepción de Tortajada et al., ningún autor había utilizado metodologías de Aprendizaje Automático y Reconocimiento de Patrones para esta tarea. Sin embargo, los buenos resultados presentados en ese estudio hacen pensar que utilizar estos métodos puede ayudar en el proceso de toma de decisiones del personal clínico, reduciendo así su carga de trabajo. También se concluye que ayudarían en la prevención de la enfermedad si se provee a las madres que han dado a luz recientemente de un test que les ofrezca un resultado mínimamente fiable sobre su riesgo de padecer DPP y que puedan realizar ellas personalmente.

A partir de datos clínicos y cuestionarios a pacientes, fue posible desarrollar una serie de nuevos modelos de clasificación. Puesto que el objetivo final de este trabajo era la creación de una aplicación de ayuda al diagnóstico y detección precoz de la DPP, orientada a personal clínico y a madres, para cada tipo de clasificador existieron dos experimentaciones distintas. Estas experimentaciones consistieron en utilizar sólo las variables que el usuario final iba a poder introducir en la entrada de la aplicación, para el entrenamiento de los modelos de clasificación. Así pues, se dividió la experimentación durante todo el trabajo en lo que llamamos modelo MÉDICO y modelo PACIENTE.

Tras el análisis de la información y la aplicación de una metodología de minería de datos para el desarrollo, validación y evaluación de los diferentes modelos de clasificación, se demuestra que las RNA presentan el mejor rendimiento en ambas experimentaciones MÉDICO y PACIENTE. Aun así, Naïve Bayes, Regresión Logística y SVM ofrecen un buen balance entre sensibilidad y especificidad.

Como se supuso desde un principio, y puesto que en el modelo MÉDICO se utilizaron 15 variables independientes para el entrenamiento, mientras que en el modelo PACIENTE sólo 10, el rendimiento en el primero fue mejor, manteniéndose un buen compromiso entre sensibilidad y especificidad en ambos.

Una comparación de los resultados de nuestros mejores clasificadores en ambas versiones de la experimentación, frente a los resultados de la mejor RNA que consiguieron Tortajada *et al.* en su estudio muestra que nuestro modelo MÉDICO mejora algunas estadísticas de la clasificación. A pesar de contar con una variable independiente menos, se consigue aumentar ligeramente el AUC, la especificidad y la tasa de aciertos. Sin embargo, los valores de G no mejoraron debido a que en nuestro caso obteníamos una sensibilidad inferior en 0.06. En todo caso, podemos destacar que se ha conseguido desarrollar un modelo con un rendimiento muy semejante, pero que necesita una variable menos de entrada.

Cabe destacar también que se consiguió una elevada especificidad de 0.81 en el modelo PACIENTE. De esta manera se espera un fallo mínimo al ofrecer un resultado negativo a las madres sanas que utilicen nuestra aplicación y una tasa aceptable de fallo al dar resultados positivos.

El siguiente objetivo que alcanzamos fue el diseño e implementación de una aplicación para dispositivos móviles que integra los mejores modelos para la predicción de la DPP conseguidos en este trabajo, en un Sistema de Ayuda a la Decisión Médica. Dicha aplicación contempla dos tipos distintos de usuarios finales: madres que hayan dado a luz recientemente y personal clínico especializado.

En su versión para madres, la aplicación realiza un cuestionario preguntando sólo por las variables que estas puedan comprender, junto con todas las preguntas de la Escala de Depresión Postparto de Edimburgo (test EPDS) para calcular su puntuación. Una vez la madre ha respondido a todas las cuestiones, la aplicación le muestra de manera comprensible el resultado de la clasificación mediante el modelo PACIENTE.

En su versión para personal clínico, la aplicación realiza un cuestionario con un lenguaje más técnico, preguntando por todas las variables independientes que se contemplaron en el modelo MÉDICO. Es en base a ese modelo sobre el que se le indica el resultado de la clasificación de la paciente dentro o fuera del grupo de riesgo de DPP.

Nuestra aplicación consigue aproximarse a la idea de un sistema de cribado eficaz y coste-eficiente ya que permite que, a través de las respuestas a unas simples preguntas, se detecten posibles casos de DPP que de otra manera quedarían sin diagnóstico.

Se ha conseguido pues, poder poner en manos tanto de personal clínico como de pacientes una herramienta que ayude a prevenir la enfermedad y a detectar la población de riesgo.

En cualquier caso, la aplicación no pretende ser un dispositivo médico, sino servir de orientación o como *screening*. El diagnóstico final siempre deberá hacerlo un psiquiatra usando el DIGS.

7. Propuesta de Actividades

A la finalización de este trabajo, la aplicación que se ha presentado es totalmente funcional, es decir, realiza todas las tareas descritas correctamente. Sin embargo, faltaría validar su usabilidad tanto con madres que acaban de dar a luz como con personal clínico. En resumen, ponerla en manos de estas personas para que pudieran dar su opinión sobre el funcionamiento, y mejorar aquellas cosas que pudieran no gustar o no entender sus potenciales usuarios. Es por ello que, de momento no se ha liberado la aplicación ni en Internet ni en la tienda de aplicaciones de *Google Play Store*.

Una vez hecho esto, sería muy interesante abarcar todos los posibles usuarios implementando la misma aplicación en las otras dos plataformas móviles más conocidas. A día de hoy, estas son las referentes a dispositivos *iPhone* y *iPad* de *Apple*, y a los dispositivos que funcionan bajo el sistema operativo *Windows Mobile* de *Microsoft*.

Además de su usabilidad, también sería posible evaluar su eficacia en la prevención de la DPP. Bajo un estudio clínico controlado, se propone hacer utilizar a un grupo de madres, después de la primera semana tras el parto, la aplicación. También se debería contar con otro grupo de las mismas características a las que no se les mencionara dicha aplicación. Pasados unos meses, sería muy interesante comprobar si existieron diferencias significativas entre las madres que usaron *eDPP Predictor* y las que no. De esta forma, podríamos comprobar si nuestra aplicación realmente funciona en su cometido.

Otro aspecto crucial que presentan las aplicaciones en dispositivos móviles es el *feedback* que puede existir con el usuario final. Para mejorar nuestros clasificadores podríamos pedir a las madres que utilizaron la aplicación, después de unos meses, que nos indiquen si han padecido o no finalmente una DPP. Esto aumentaría de manera notable el volumen de información de nuestra base de datos, y por lo tanto podríamos reentrenar nuestros modelos o conseguir unos nuevos con esta nueva información con el fin de mejorar los resultados que ya tenemos. No ocultamos el riesgo de que esta información no sea validada por personal clínico.

Otras mejoras sobre la versión orientada a personal clínico, pasan sin duda por hacer que la información de las pacientes pueda ser almacenadas en una base de datos interna para comodidad y posible recuperación de datos del médico. De este modo podría ejercer un mejor control de sus pacientes, además de poder proporcionarnos un *feedback* con datos clínicamente validados si así lo desea. En ambos casos, este *feedback* se debería realizar teniendo especial cuidado en el cumplimiento de la Ley Orgánica de Protección de Datos (LOPD) y del consentimiento informado.

Respecto a los modelos de clasificación que no se emplearon durante este estudio, sería interesante comprobar el rendimiento en este problema de las Redes Bayesianas, puesto que también pueden hacer frente a pérdidas de información, encontrar dependencias probabilísticas y presentar una buenas prestaciones [70].

8. Bibliografía

1. International Classification of Diseases. ICD-10 Version:2010. [Online] http://apps.who.int/classifications/icd10/browse/2010/en#/F53.0.

2. **Association, American Psychiatric.** *Diagnostic and Statistical Manual of Mental Disorders. 4a ed. revised (DSM-IV-TR).* Washington : American Psychiatric Press, 2000.

3. *Depresión posparto.* **Agnès Arbat, Imma Danés.** Medicina clínica, ISSN 0025-7753, Vol. 121, Nº. 17, 2003 , págs. 673-675.

4. **Vega, Montserrat García.** *Trastornos del estado de ánimo en el puerperio: Factores psicosociales predisponentes.* Madrid : s.n., 2010. ISBN: 978-84-693-9426-7.

5. *Women at risk for postpartum-onset major depression.* **Zachary N. Stowe, Charles B. Nemeroff.** 1995, American Journal of Obstetrics and Gynecology, Vol. 173, pp. 639–645.

6. **Diana Marcela Peña, José Manuel Calvo.** Aspectos clínicos de la depresión postparto. *Obstetricia integral Siglo XXI.* 2011.

7. *Postpartum mood disorders.* **Seyfried L, Marcus M.** 2003, International Review of Psychiatry, Vol. 15, pp. 231-242.

8. *Treatment of Mood Disorders During Pregnancy and Postpartum.* **Lee S. Cohen, Betty Wang, Ruta Nonacs, Adele C. Viguera, Elizabeth L. Lemon, Marlene P. Freeman.** 2, 2010, Psychiatric Clinics of North America, Vol. 33, pp. 273-293.

9. *Prediction, detection and treatment of postnatal depression.* **Cooper PJ, Murray L.** 1997, Archives of Disease in Childhood, Vol. 77, pp. 97-99.

10. **Stewart, D.E., Robertson, E., Dennis, C-L., Grace, S.L., & Wallington, T.** *Postpartum depression: Literature review of risk factors and interventions.* Toronto : s.n., 2003.

11. *Prevalencia de la depresión posparto en las madres españolas: comparación de la estimación.* **Carlos Ascaso Terréna, Lluïsa Garcia Esteveb, Puri Navarrob, Jaume Aguadoa.** Med Clin (Barc). 2003;120:326-9. - vol.120 núm 09 .

12. *Prediction of Postpartum Depression Using Multilayer Perceptrons and Pruning.* **Tortajada S., Garcia-Gomez J. M., et al.** 3, 2009, Methods of Information in Medicine, Vol. 48, pp. 291-298.

13. *Antenatal prediction of postpartum depression with blood DNA methylation biomarkers.* **Guintivano J, Arad M, Gould TD, Payne JL, Kaminsky ZA.** 2013, Mol Psychiatry. DOI: 10.1038/mp.2013.62.

14. *Rates and risk of postpartum depression--a meta-analysis.* **O'Hara MW, Swain AM.** 1, 1996, Int Rev Psychiatry, Vol. 8, pp. 37-54. DOI: 10.3109/09540269609037816.

15. *Development of the 10-item Edinburgh Postnatal Depression Scale.* **Cox JL, Holden JM, Sagovsky R.** 150, 1987, Br J Psychiatry, pp. 782-6.

16. **Department of Health, Government of Western Australia.** *Depression Scale (EPDS): Translated versions – validated.* [ed.] Western Australia: State Perinatal Mental Health Reference Group. Edinburgh Postnatal Perth. 2006.

17. *Implications of timing of maternal depressive symptoms for early cognitive and language development.* **Sohr-Preston SL, Scaramella LV.** 2006, Clin Child Fam Psychol Rev., Vol. 9, pp. 65-83.

18. *Suicide during pregnancy and in the first postnatal year.* **Appleby, L.** 302, 1991, BMJ, pp. 137-140.

19. *Suicide: the leading cause of maternal death.* **Oates, M.** 183, 2003, Br J Psychiatry, pp. 279-281.

20. *Review of screening instruments for postpartum depression.* **Boyd RC, Le HN.** 2005, Arch Womens Ment Health, Vol. 8, pp. 141-153.

21. *Evidence-based medicine: a unified approach.* **Eddy, DM.** 1, 2005, Health Aff (Millwood), Vol. 24, pp. 9-17.

22. **Grain, H.** *Guide to the principles and desirable features of clinical decision support.* s.l. : Standards Australia, 2007.

23. **Bishop, CM.** *Patern Recognition and Machine Learning.* s.l. : Springer, 2006. ISBN: 0-38-731073-8.

24. **C.-L., Dennis.** Detection, prevention, and treatment of postpartum depression. [book auth.] Robertson E., Dennis C.-L., Grace, S.L., Wallington, T. Stewart D.E. *Postpartum depression: Literature review of risk factors and interventions.* 2003.

25. *Prediction of depression in the postpartum period: a longitudinal follow-up study in high-risk and low-risk women.* **Gerda J.M. Verkerka, Victor J.M. Popa, Maarten J.M. Van Sonb, Guus L. Van Hecka.** 77, 2003, Journal of Affective Disorders, pp. 159-166.

26. *Comparative analysis of the performance of the Postpartum Depression Screening Scale with two other depression instruments.* **Beck CT, Gable RK.** 4, 2001, Nursing Research, Vol. 50, pp. 242-50.

27. *Identifying Mothers with Postpartum Depression Early: Integrating Perinatal Mental Health Care into the Obstetric Setting.* **Helen Chen, Jemie Wang, Ying Chia Ch'ng, Roshayati Mingoo, Theresa Lee, and Julia Ong.** 309189, 2011, ISRN Obstetrics and Gynecology, Vol. 2011. DOI:10.5402/2011/309189.

28. *Mood changes after delivery: role of the serotonin transporter gene.* **J. Sanjuan, R. Martin-Santos, L. Garcia-Esteve et al.** 2008, The British Journal of Psychiatry, Vol. 193, pp. 383-388. DOI: 10.1192/bjp.bp.107.045427 .

29. *Validation of the Edinburgh Postnatal Depression Scale (EPDS) in Spanish mothers.* **García-Esteve L, Ascaso L, Ojuel J, Navarro P.** 2003, Journal of Affective Disorders, Vol. 75, pp. 71-76.

30. *Diagnostic interview for genetic studies and training.* **Nurnberger JI, Blehar MC, Kaufmann C, York-Cooler C, Simpson S, Harkavy-Friedman J, et al.** 1994, Archives of Genetic Psychiatry, Vol. 51, pp. 849-859.

31. *Diagnostic Interview for Genetic Studies (DIGS): Inter-rater and test-retest reliability and validity in a Spanish population.* **Roca M, Martin-Santos R, Saiz J, Obiols J, Serrano MJ, Torrens M, et al.** 2007, European Psychiatry, Vol. 22, pp. 44-48.

32. **Eysenck HJ, Eysenck SBG.** *The Eysenck Personality Inventory.* London : University of London Press, 1964.

33. *A psychometric analysis of the revised Eysenck Personality Questionnaire short scale.* **Aluja A, García O, García LF.** 2003, Personality and Individual Differences, Vol. 35, pp. 449-460.

34. **Ministerio de Salud, Gobierno de Chile.** *Guía Clínica Depresión en personas de 15 años y más.* Santiago : MINSAL, 2013.

35. *Methodological aspects of life events research.* **ES., Paykel.** 1983, Journal of Psychosomatic Research, Vol. 27, pp. 341-352.

36. *Association of a triallelic serotonin transporter gene promoter region (5-HTTLPR) polymorphism with stressful life events and severity of depression.* **Zalsman G, Huang YY, Oquendo MA, Burke AK, Hu XZ, Brent DA, et al.** 2006, American Journal of Psychiatry, Vol. 163, pp. 1588-1593.

37. *Validity and reliability of the Duke-UNC-11 questionnaire of functional social support.* **Bellón JA, Delgado A, Luna JD, Lardelli P.** 1996, Atención Primaria, Vol. 18, pp. 158-163.

38. *Serotonin transporter promoter and intron 2 polymorphisms: relationship between allelic variants and gene expression.* **Hranilovic D, Stefulj J, Schwab S, Borrmann-Hassenbach M, Albus M, Jernej B, et al.** 2004, Biological Psychiatry, Vol. 55, pp. 1090-1094.

39. **Velert, Salvador Tortajada, dir., García Gómez JM and dir., Robles Viejo M.** *Incremental Learning approaches to Biomedical decision problems.* Valencia : s.n., 2012.

40. **George E P Box, Norman R Draper.** *Empirical Model-Building and Response Surfaces.* s.l. : Wiley, 1987. p. 424. ISBN 0-471-81033-9.

41. **Breiman L, Friedman J, Olshen R, Stone C.** *Classification and Regression Trees.* Boca Raton : CRC Press, 1984.

42. **Dasarathy, Belur V.** *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.* 1991. ISBN 0-8186-8930-7.

43. **Mitchell, T.** *Machine Learning.* s.l. : McGraw Hill, 1997.

44. **David G. Kleinbaum, Mitchel Klein.** *Logistic Regression. A Self-Learning Text.* s.l. : Springer, 2010. ISBN: 978-1-4419-1741-6.

45. **Sayad, Dr Saed.** An Introduction to Data Mining. *Logistic Regression.* [Online] http://www.saedsayad.com/logistic_regression.htm.

46. **Hosmer DW, Lemeshow S.** *Applied logistic regression.* s.l. : Wiley-Interscience, 2000.

47. **Shawe-Taylor, John and Cristianini, Nello.** *An introduction to Support Vector Machines and other kernel-based learning methods.* s.l. : Cambridge University Press, 2000. ISBN: 0 521 78019 5.

48. **Sayad, Dr Saed.** An Introduction to Data Mining. *Support Vector Machine - Classification (SVM).* [Online] http://www.saedsayad.com/support_vector_machine.htm.

49. *Applying machine learning to software fault-proneness prediction.* **Gondra, Iker.** 2008, The Journal of Systems and Software, Vol. 81, pp. 186-195.

50. *The Perceptron: a probabilistic model for information storage and organization in the brain.* **F, Rosenblatt.** 6, 1958, Psychological Review, Vol. 65, pp. 386-408.

51. **CM, Bishop.** *Neural Networks for Pattern Recognition.* Oxford : Clarendon Press, 1995.

52. *Pruebas diagnósticas: Sensibilidad y especificidad.* **Pita Fernández S, Pértegas Díaz S.** A Coruña : s.n., 2003, Cad Aten Primaria, Vol. 10, pp. 120-124.

53. *The class imbalance problem: a systematic study.* **Japkowicz N, Stephen S.** 5, 2002, Intelligent data analysis Journal, Vol. 6, pp. 429-449.

54. *An introduction to ROC analysis.* **T, Fawcett.** 8, 2006, Pattern Recognition Letters, Vol. 27, pp. 861-874.

55. *Uso de curvas ROC en investigación clínica. Aspectos teórico-prácticos.* **J Cerda, L Cifuentes.** 2, Revista chilena de infectología, Vol. 29, pp. 138-141.

56. Curva ROC. [Online] http://es.wikipedia.org/wiki/Curva_ROC.

57. **Pyle, Dorian.** *Data Preparation for Data Mining.* s.l. : Morgan Kaufmann Publishers, 1999. ISBN-10: 1558605290.

58. **Fernando Medina, Marco Galván.** *Imputación de datos: teoría y práctica.* [ed.] División de Estadística y Proyecciones Económicas Naciones Unidas. Santiago de Chile : s.n., 2007. ISBN: 978-92-1-323101-2.

59. *On the Meaning and Use of Kurtosis.* **DeCarlo, Lawrence T.** 3, 1997, Psychological Methods, Vol. 2, pp. 292-307.

60. *The Kolmogorov-Smirnov Test for Goodness of Fit.* **FJ, Massey.** 253, 1951, Journal of American Statistical Association, Vol. 46, pp. 68-78.

61. *Table of Percentage Points of Kolmogorov Statistics.* **LH, Miller.** 273, 1956, Journal of the American Statistical Association, Vol. 51, pp. 111-121.

62. **Gibbons JD, Chakraborti S.** *Nonparametric Statistical Inference, 5th Ed.* Boca Raton : Chapman & Hall/CRC Press, Taylor & Francis Group, 2011.

63. **Hollander M, Wolfe DA.** *Nonparametric Statistical Methods.* s.l. : John Wiley & Sons, 1999.

64. **AJ, Dobson.** *An introduction to Generalized Linear Models.* New York : Chapman & Hall, 1990.

65. **Masters, T.** Practical Neural Networks recipes in C++. s.l. : Academic Press, Inc., 1993, pp. 173-180.

66. Matlab R2013a Online Documentation. *Levenberg-Marquardt backpropagation.* [Online] http://www.mathworks.es/es/help/nnet/ref/trainlm.html.

67. Matlab R2013a Online Documentation. *Hyperbolic tangent sigmoid transfer function.* [Online] http://www.mathworks.es/es/help/nnet/ref/tansig.html.

68. **Gironés, Jesús Tomás.** *El gran libro de Android.* s.l. : Marcombo SA, 2012. ISBN 978-84-267-1832-7.

69. Statista - The Statistics Portal. *Number of daily activations of Android devices from August 2010 to March 2013.* [Online] http://www.statista.com/statistics/219554/daily-activations-of-android-devices/.

70. *Accuracy in the diagnostic prediction of acute appendicitis based on the bayesian network model.* **Sakai S, Kobayashi K, Nakamura J, Toyabe S, Akazawa K.** 2007, Methods Inf Med, Vol. 46, pp. 723-726.