Universidad del País Vasco
Euskal Herriko Unibertsitatea

Máster Universitario en

Ingeniería **C**omputacional y **S**istemas **I**nteligentes

Konputazio Zientziak eta Adimen Artifiziala Saila –
Departamento de Ciencias de la Computación e Inteligencia Artificial

K I S A

I C S I

Tesis de Máster

Caracterizando un ciclo de reproducción asistida exitoso mediante clasificación supervisada y positive unlabelled

Mikel Sánchez Corujo

Tutores

Jerónimo Hernández González
Departamento de Ciencia de la Computación e Inteligencia Artificial
Facultad de Informática

Iñaki Inza Cano
Departamento de Ciencia de la Computación e Inteligencia Artificial
Facultad de Informática

informatika fakultatea
facultad de informática

KZAA /CCIA

Septiembre 2018

Índice general

Capítulo 1

Introducción

En medicina, el problema de las tecnologías de reproducción asistida (del inglés, *Assisted Reproductive Technologies*, ART) ha recibido una atención considerable. Este problema consiste en solventar la dificultad de inducir un embarazo sin aumentar las probabilidades del intrínsecamente arriesgado embarazo múltiple. Durante todo el procedimiento se tienen que tomar muchas decisiones médicas y, por consiguiente, el objetivo de las líneas de investigación actuales es aumentar el conocimiento sobre el problema para apoyar las decisiones de los médicos. Con este objetivo, se han aplicado diferentes técnicas de inteligencia artificial y de aprendizaje automático al problema de las ART.

En colaboración con la Unidad de Reproducción Asistida del Hospital de Donostia, se propone una solución para el problema del ART. El objetivo principal es obtener evidencias sobre la relevancia de los datos recopilados y su uso potencial para mejorar la tasa de embarazos. Se proponen dos enfoques diferentes que proporcionan información valiosa para resolver parcialmente el problema del ART: el primero de ellos, basado en la clasificación supervisada clásica; y un segundo que se configura como un problema de clasificación débilmente supervisada. Al contrario de la práctica habitual, donde se suelen descartar los embriones de destino desconocido, las técnicas débilmente supervisadas consideran incluso las instancias de embriones/ciclos cuyo destino no se puede establecer con certeza. En los experimentos realizados, los dos enfoques han arrojado interesantes resultados.

1.1. Motivación

Desde que el 25 de julio de 1978 nació Louise Joy Brown, el primer nacimiento obtenido por Fecundación in Vitro (FIV), el problema de las tecnologías de reproducción asistida ha recibido considerable atención en el mundo de la medicina. Un campo para el que, en general, se considera que todavía hay espacio para avances [1]. Las tecnologías de reproducción asistida son

un conjunto de técnicas médicas invasivas que intentan inducir un embarazo. Éstas incluyen tratamientos de fertilidad en los que se manipulan óvulos o embriones en un laboratorio. Aunque la mayor parte de los procedimientos de ART realizados son por FIV, también forman parte de estas tecnologías otras técnicas como la inyección intracitoplasmática de espermatozoides (del inglés, *IntraCytoplasmic Sperm Injection*, ICSI), la transferencia intrafalopiana de gametos o la transferencia intrafalopiana de cigotos. Debido a que un procedimiento de ART consta de varios pasos en un intervalo de aproximadamente 2-3 semanas, a cada tratamiento de ART se le denomina *ciclo*.

Un ciclo de ART generalmente comienza con un tratamiento farmacológico con el fin de provocar la estimulación ovárica de la mujer para inducir el desarrollo de múltiples folículos con una gran cantidad de ovocitos [2]. Si éstos se producen, el ciclo progresa a la etapa de recuperación de los mismos, que implica la extirpación quirúrgica de los óvulos de los ovarios. Tras recuperarlos, los ovocitos maduros se fecundan en el laboratorio con el esperma del hombre durante el procedimiento de FIV y los embriones resultantes se cultivan durante varios días. Si tiene éxito, llega el momento de tomar la decisión de elegir el/los embrión/es que parecen más viables para inducir un embarazo. Los médicos serán los encargados de seleccionar los mejores embriones para desarrollarse e implantarse basándose en diferentes variables, como su morfología. Después de la transferencia al útero, si un embrión se implanta, se diagnostica un embarazo clínico por la presencia de un saco gestacional detectable por ultrasonido. Este es un proceso natural que no puede ser supervisado por el especialista, y que determina el éxito de un ciclo de ART. La implantación de al menos uno de los embriones transferidos lleva el ciclo a un embarazo.

El punto crítico de este proceso es la elección de los embriones a transferir por parte de los médicos. Para aumentar el conocimiento de los expertos a la hora de elegir los mejores embriones, en esta memoria se estudiará la viabilidad del ciclo intentando aislar la influencia del embrión para identificar las características de los ciclos exitosos.

1.2. Explicación del problema

En las últimas décadas, se ha discutido mucho sobre las características que determinan el éxito de un ciclo de entre las recogidos en las diferentes etapas del procedimiento de ART. Entre estas características se consideran variables que describen el ciclo como las referidas a la evaluación femenina y masculina o la estimulación ovárica y, por otra parte, aquéllas que describen cada ovocito/embrión. Como recoge Hernández [3], en sus revisiones exhaustivas, Achache y Revel [4] y Ebner et al. [5] recopilaron y discutieron un amplio conjunto de variables que se han considerado para evaluar la calidad de los ciclos y los ovocitos/embriones.

El objetivo de este estudio es mejorar la tasa de embarazo de las ART. Diferentes estudios han demostrado que el número de embriones a transferir se correlaciona positivamente con la probabilidad de embarazo [6]. Como consecuencia, en los datos con los que se ha trabajado se observa que en alrededor del 87 % de los ciclos se ha producido una transferencia múltiple, es decir, se ha transferido más de un embrión en el mismo ciclo. Sin embargo, la multi-transferencia de embriones en un mismo ciclo puede dar cabida a la implantación conjunta de varios de ellos, lo que eventualmente conduce a un embarazo múltiple, que es ampliamente considerado de riesgo para la mujer y el desarrollo de los fetos [6]. Para reducir los riesgos que puede conllevar un embarazo múltiple, se han establecido limites legales en el número máximo de embriones a transferir, que en el caso de España está limitado a un máximo de 3 embriones por ciclo.

Dado que el procedimiento clínico de ART generalmente produce embriones en exceso, los médicos deben seleccionar aquéllos más adecuados para transferir y poder concluir el ciclo del ART con éxito al producirse un embarazo. Los embriones a transferir han de ser cuidadosamente seleccionados ya que la transferencia de los embriones de mala calidad es una contribución importante al fracaso de las ART [5]. En la literatura relacionada, todas estas consideraciones han originado una interesante discusión sobre la posibilidad de seleccionar y transferir un conjunto de embriones prometedores, o sólo uno, que conducirá a un embarazo simple [6].

Tradicionalmente, el estudio de la reproducción asistida tiende a modelar separadamente la posibilidad de éxito del ciclo y la de los embriones obtenidos/empleados. Sin embargo, están relacionados: sólo se puede estar seguro de que un ciclo es viable al llevarlo a cabo y obtener un embarazo, lo que implica necesariamente el uso de un embrión viable.

Dadas unas bases de datos reales facilitadas por la Unidad de Reproducción Asistida del Hospital de Donostia, con la intención de modelar y aislar la influencia de los embriones en el aprendizaje de un modelo, se proponen dos enfoques diferentes que brindan información valiosa para resolver parcialmente el problema: en el primero de ellos, se tratará de construir un modelo de clasificación supervisada para la predicción del embarazo; y, en el segundo, se tratará de modelar la configuración de un ciclo prometedor desde una perspectiva de clasificación débilmente supervisada.

1.2.1. Clasificación supervisada: Predicción de Embarazo

El primero de los enfoques del estudio consiste en intentar predecir si un nuevo ciclo de ART culminará en un embarazo. Para tratar de dar respuesta a ese problema se aplicarán técnicas de clasificación supervisada. El concepto de clasificación supervisada [7, 8] se refiere a la inducción de clasificadores a partir de un conjunto de instancias etiquetadas. Se puede ver como un intento de emular el aprendizaje humano a partir de la experiencia. De modo que

los algoritmos supervisados intentan imitar el cerebro humano creando un modelo del problema en cuestión, a partir de la experiencia recogida en el conjunto de datos.

En este enfoque se construye un modelo de clasificación que, utilizando los datos recolectados de ciclos anteriores, predice si un nuevo ciclo terminará en un embarazo. Cada ejemplo de entrenamiento representa un ciclo de ART con su categoría real. Un buen clasificador vaticinará con precisión la etiqueta de clase de nuevos ciclos sin etiquetar.

Los ciclos se etiquetan con acierto una vez pasadas cinco semanas de la transferencia de los embriones, cuando la existencia de un embarazo evolutivo se puede evaluar mediante el uso de técnicas de ultrasonido. Diversos autores han utilizado este enfoque para evaluar la relevancia de un conjunto reducido de variables supuestamente determinantes (por ejemplo, la edad de la mujer, el tiempo de esterilidad o el número de ciclos anteriores) [9]. Por el contrario, todas las características recogidas por los médicos se han considerado inicialmente como variables predictivas (Tabla 1.1) [3].

Aunque las técnicas de aprendizaje calibran automáticamente la contribución de cada característica, el uso de un conjunto más grande de variables generalmente introduce variables irrelevantes y/o redundantes, lo que puede ser perjudicial para el rendimiento del clasificador [10]. Por lo tanto, se aplican técnicas de selección de subconjuntos de características (del inglés, *Feature Subset Selection*, FSS) [11] para identificar automáticamente las variables predictivas relevantes y descartar aquéllas que no aportan información valiosa.

En este trabajo se utiliza el modelo naive Bayes para construir la función de clasificación que, basándose en ciclos reales facilitados (junto a su etiqueta de clase correspondiente), reproduce el comportamiento de categorización de un ciclo futuro.

El objetivo de estos clasificadores es predecir la etiqueta de clase de ciclos nuevos con cierto grado de precisión. Una identificación precisa de los ciclos que terminarán en un embarazo se reflejaría seguramente en una mejora significativa del rendimiento de las ART. Para evaluar la bondad del modelo aprendido se aplicarán diferentes métricas de clasificación supervisada. Además, también se estudiarán los resultados obtenidos al agregar a las variables de ciclos las variables de los embriones para valorar la influencia de los mismos en el proceso del tratamiento.

1.2.2. Clasificación PU: Predicción de viabilidad de los embriones

El segundo enfoque tratará de identificar la configuración de un ciclo prometedor. Los clasificadores construidos en este enfoque no se pueden utilizar para predecir embarazos, sino que identificarán ciclos de buen pronóstico que, en el mejor escenario (es decir, se lleva a cabo con embriones promete-

dores sin ocurrencia de un MIF[1]), terminarían en un embarazo. En su tésis doctoral, Hernández [3] denomina este enfoque como la predicción *embarazable*. En este planteamiento, los ejemplos de entrenamiento están descritos por todas las características del ciclo de la Tabla 1.1; excepto de aquéllas que describen el proceso de transferencia: *No.Transf.Emb* y *SelectiveTransf.*

El conjunto de datos de entrenamiento se divide en dos subconjuntos claramente separables. Un subconjunto de ejemplos positivos, que se compone de todos los ciclos que terminaron en un embarazo. Cualquier ciclo que termina en embarazo implica la existencia de un ciclo *"bueno"* (*embarazable*). Sin embargo, que un ciclo fracase no siempre implica una mala configuración del mismo. Por ese motivo, un ciclo que no ha culminado en un embarazo no puede ser considerado como un ejemplo negativo de este enfoque, puesto que puede ser debido a la transferencia de embriones de mala calidad o a la ocurrencia de un MIF. Como las técnicas médicas actuales no pueden determinar qué causa específica es responsable del fracaso de cada ciclo, el segundo subconjunto está formado por esos ejemplos que han fracasado y que se anotan como no etiquetados. Esto se modela a través de otro paradigma de clasificación débilmente supervisado, el aprendizaje positivo no etiquetado (del inglés, *Positive Unlabelled*, PU) [13], donde un clasificador binario se aprende de un conjunto de datos que sólo contiene instancias positivas y la mayoría de los ejemplos no tienen etiqueta. Las técnicas PU permiten aprender de este tipo de datos, teniendo en cuenta que en el subconjunto no etiquetado podría haber tanto ejemplos positivos como negativos.

Este tipo de problema es bastante frecuente en el marco de recuperación de información. Se tiene un gran conjunto de instancias no etiquetadas y una lista de ejemplos etiquetados, y lo que se busca es recuperar, desde el gran conjunto no etiquetado, aquellas instancias que son similares a las etiquetadas positivamente. En general, la falta de información, tiempo o dinero puede hacer que sea difícil obtener un conjunto de ejemplos negativos que representen todas las posibles clases de instancias sin interés, por lo tanto, la clasificación PU es una alternativa para aprender clasificadores sin ejemplos negativos.

Dado que en este problema se carece de instancias de clase negativa, no se puede trabajar con los mismos clasificadores del aprendizaje supervisado. Sin embargo, se utilizarán dos técnicas diferentes. En la primera se presentará un algoritmo de dos pasos en el que, en un primer momento, se obtiene una muestra de instancias que podría ser considerada negativa; y, en el segundo, se trata de predecir la clase de las instancias no etiquetadas a partir de un modelo construido con el conjunto positivo y el negativo obtenido. Esta técnica es el algoritmo *spy-EM* [14]. Por otra parte, se aplicará una

[1]Con el conocimiento que se tiene en la actualidad, cuando fracasa un ciclo de ART para un embrión con características que lo hacen parecer prometedor se dice que ha ocurrido error de implantación incomprendido (del inglés, *Misunderstood Implantation Failure*, MIF) [12].

adaptación del modelo naive Bayes al cuadro de clasificación PU.

Además, dado que no todas las variables de la base de datos tienen por qué ser igual de relevantes, para construir el clasificador, se introducirá una técnica para reducir la dimensión del problema de clasificación.

Por último, se intentará aprovechar la información disponible sobre las características de los embriones empleados en cada ciclo, para crear nuevas variables que se puedan utilizar para el estudio de los ciclos de ART. Dado que en cada ciclo hay varios embriones involucrados, para comprobar la influencia o no de las características relativas a éstos, se van a construir variables que trasladen la información de los N embriones implicados en cada ciclo a las variables del ciclo. Debido a esa relación 1-a-N, las variables que se creen serán variables agregadas a las características del ciclo referidas a los embriones. A partir de las variables de embriones de la Tabla 1.2, se han creado las variables agregadas descritas en la Tabla 1.3.

	Variable	Valores	Descripción
Ciclo	Indicación	endometriosis, fracaso inseminación artificial, factor tubárico, masculino, mixto, otros, desconocido	Indicación del ciclo
Femeninas	Tiempo Esteril	$(0,3]$, $(3,5]$, $(5,inf)$	Tiempo desde que se detectó la esterilidad
	Edad	$(0,30]$, $(30,35]$, $(35,inf)$	Edad
	IMC	$(0,20]$, $(20,25]$, $(25,inf)$	Índice de masa corporal
	Emb. previos	Sí, No	¿Ha estado embarazada previamente?
	Abo. previos	Sí, No	¿Ha sufrido algún aborto previamente?
	Ciclos previos	0, 1, +2	Número de ciclos de ART a los que se ha sometido previamente
	FSH	$(0,5]$, $(5,8]$, $(8,inf)$	Cantidad de hormona foliculoestimulante
	AMH	$(0,2]$, $(2,inf)$	Cantidad de hormona antimulleriana
	Fol. Antral	$(0,5]$, $(5,10]$, $(10,inf)$	Cantidad de folículos antrales
	E2	$(0,2000]$, $(2000,3000]$, $(3000,inf)$	Cantidad de estradiol
	P4	$(0,0.5]$, $(0.5,1.25]$, $(1.25,inf)$	Cantidad de progesterona
	lEnd	$(0,9]$, $(9,12]$, $(12,inf)$	Grosor endometrial
Masc.	Cal. Semen	A, N, O, OA, OAT	Calidad del semen
	REM	$(0,5]$, $(5,15]$, $(15,inf)$	Recuento de espermatozoides móviles
Estimulación	Protocolo	PC+Agon, PC+Ant, PL	Protocolo de estimulación
	Estimulación	FSH+Lhrec, FSHrec, FSHrec+hMG, FSHur, FSHur+hMG, hMG	Tratamiento de estimulación
	dEst	$(0,9]$, $(9,11]$, $(11,inf)$	Días de tratamiento de estimulación
	unidFSH	$(0,2000]$, $(2000,3000]$, $(3000,inf)$	Unidades de FSH
	unidLH	0, $(0,1500]$, $(1500,inf)$	Unidades de LH
Embriones	nOvocit	$(0,5]$, $(5,10]$, $(10,inf)$	Número de ovocitos recuperados
	MII	$(0,5]$, $(5,10]$, $(10,inf)$	Número de ovocitos maduros (en estado meiosis MII)
	nEmbObten	$(0,3]$, $(3,8]$, $(8,inf)$	Número de embriones obtenidos
	TasaFertil	$(0,0.5]$, $(0.5,0.75]$, $(0.75,1]$	Tasa de fertilidad (nEmbObten / MII)
	nEmbTrans	0, 1, 2, 3	Número de embriones transferidos
	transSelect	Sí, No	¿Se seleccionaron los embriones transferidos? (nEmbObten > nEmbTrans)
	transDia	D+2, D+3, Blasto	Día en el que se produjo la transferencia
	transDificult	fácil, difícil	Nivel de dificultad de la transferencia
	transSangre	Sí, No	¿Se ha sangrado en la transferencia?
	transIntento	1, 2	Número de intentos de la transferencia
	calEmbTrans	A, B, C, D, ...	Calidad de los embriones transferidos
	vitrific	Sí, No	¿Se han vitrificado los embriones sobrantes?
Salida	Embarazo	Sí, No	¿Se ha producido un embarazo?
	No.Sacs	0, 1, 2, 3	Número de sacos gestacionales

Tabla 1.1: Tabla con las variables de ciclos [3].

	Variable	Valores	Descripción
Oocito	Vac	Sí, No	Presencia de vacuolas
	SER	Sí, No	Presencia de clusters de retículo endoplasmático liso
	PVS	Normal, Aumentada	Descripción del espacio perivitelino
	PB	Normal, Anormal	Descripción del primer cuerpo polar
	Técnica	IVF, ICSI	Técnica de Fertilización
D+1	PB.1	1, 2, 3+	Número de cuerpos polares
	Z	Z1, Z2, Z3, Z4	Grado pronuclear de Scott [15]
D+2	nCel.2	$\{4\}$, $\{2;5\}$, {otros}	Número de células
	frag.2	$[0,10]$, $(10,25)$, $(25,35]$, $(35,100)$	Porcentaje de fragmentación celular
	symmet.2	Sí, No	Simetría en los blastómeros
	PZ.2	Normal, Anormal	Presencia de anormalidades en la zona pelúcida
	vac.2	Sí, No	Presencia de vacuolas
	multiNuc.2	Sí, No	Presencia de núcleo múltiple en una célula ($no.nuclei. \geq 2$)
	Calidad.2	A, B, C, D	Grado de calidad de ASEBIR [16]
	Transferencia	Sí, No	Embrión seleccionado para transferencia
Salida	Implantado	Sí, No	Embrión implantado
	Blastocito	Sí, No	¿Alcanzó la etapa de blastocito?

Tabla 1.2: Tabla con las características recopiladas para los ovocitos / embriones. Las variables Implantado y Blastocito no siempre pueden ser anotadas por los médicos [3].

Variable	Valores	Descripción
modaCalidad	A, B, C, D	Moda del grado de calidad de ASEBIR [16]
numEmbCalA	0, 1, 2, 3+	Número de embriones de calidad A
numEmbCalB	0, 1, 2, 3+	Número de embriones de calidad B
numEmbCalC	0, 1, 2, 3+	Número de embriones de calidad C
numEmbCalD	0, 1, 2, 3+	Número de embriones de calidad D
modaZ	Z1, Z2, Z3, Z4	Moda del grado pronuclear de Scott [15]
numEmbZ1	0, 1, 2, 3+	Número de embriones de grado Z1
numEmbZ2	0, 1, 2, 3+	Número de embriones de grado Z2
numEmbZ3	0, 1, 2, 3+	Número de embriones de grado Z3
numEmbZ4	0, 1, 2, 3+	Número de embriones de grado Z4
numEmb4cels	0, 1, 2, 3+	Número de embriones con 4 células
modaNFrag	$(0, 10]$, $(10, 25]$, $(25, 35]$, $(35, inf)$	Moda del porcentaje de fragmentación celular
numEmbFragN1	0, 1, 2, 3+	Número de embriones con porcentaje de fragmentación celular $(0, 10]$
numEmbFragN2	0, 1, 2, 3+	Número de embriones con porcentaje de fragmentación celular $(10, 25]$
numEmbFragN3	0, 1, 2, 3+	Número de embriones con porcentaje de fragmentación celular $(25, 35]$
numEmbFragN4	0, 1, 2, 3+	Número de embriones con porcentaje de fragmentación celular $(35, inf)$
modaSimet	Sí, No	Moda de la simetría en los blastómeros
numEmbSimetNo	0, 1, 2, 3+	Número de embriones con asimetría en los blastómeros
numEmbSimetSi	0, 1, 2, 3+	Número de embriones con simetría en los blastómeros
numEmbVacSi	0, 1, 2, 3+	Número de embriones con presencia de vacuolas
numEmbVacNo	0, 1, 2, 3+	Número de embriones sin presencia de vacuolas
numEmbmultinucSi	0, 1, 2, 3+	Número de embriones con presencia de núcleo múltiple en una célula
numEmbmultinucNo	0, 1, 2, 3+	Número de embriones sin presencia de núcleo múltiple en una célula
modaCalTransfer	A, B, C, D	Moda del grado de calidad de ASEBIR de los embriones transferidos
modaZTransfer	Sí, No	Moda del grado pronuclear de Scott de los embriones transferidos
modaNfragTransfer	$(0, 10]$, $(10, 25]$, $(25, 35]$, $(35, inf)$	Moda del porcentaje de fragmentación celular de los embriones transferidos
modaSimetTransfer	Sí, No	Moda de la simetría en los blastómeros de los embriones transferidos
modaVacTransfer	Sí, No	Moda de presencia de vacuolas en los embriones transferidos
modaMultinucTransfer	Sí, No	Moda de presencia de núcleo múltiple en una célula de los embriones transferidos

Tabla 1.3: Tabla con las características agregadas de las variables de embriones basadas en los criterios del manual ASEBIR [16].

Capítulo 2

Problemas de clasificación

En este capítulo, se presentan los conceptos matemáticos necesarios para entender el trabajo realizado. En primer lugar, se hace una introducción general a los problemas de clasificación y la notación que se va a utilizar. Después, se explican con más detalle los dos problemas de clasificación que se han trabajado: la clasificación supervisada y la clasificación PU.

2.1. Los problemas de clasificación

El punto de partida en cualquier problema de clasificación es un conjunto de datos $\mathcal{D}$ de instancias o ejemplos. Estas instancias se representan por un vector de n variables aleatorias o características, $\boldsymbol{X} = (X_1, ..., X_n)$, que pueden ser discretas o continuas[1]. Una variable aleatoria discreta X_i puede tomar r_i valores. Las variables aleatorias se representan con letras mayúsculas y sus valores con minúsculas; mientras que los vectores están representados en negrita.

Las instancias se clasifican según el criterio representado por la variable clase $\boldsymbol{C}$ que toma r_C valores en $\Omega_{\boldsymbol{C}}$[2]. Las variables de predicción y de clase pueden tener valores perdidos que se representarán con un signo de interrogación. $\mathcal{D}$ denota un conjunto de datos de N ejemplos $\mathcal{D} = \{(\boldsymbol{x}^{(1)}, c^{(1)}), ..., (\boldsymbol{x}^{(N)}, c^{(N)})\}$.

Un problema de clasificación puede definirse como la inducción, a partir de un conjunto de datos $\mathcal{D}$, de una función de clasificación φ que, dado el vector de características de una instancia, devuelve una etiqueta de clase c.

$$\begin{aligned} \varphi : \ \Omega_{\boldsymbol{X}} &\longrightarrow \Omega_{\boldsymbol{C}} \\ \boldsymbol{x} &\longmapsto c \end{aligned}$$

[1] En este trabajo se trabaja sólo con variables discretas. Aquéllas que no lo son se han categorizado como se muestra en las Tablas 1.1 y 1.2

[2] La variable clase C es discreta. La predicción de variables continuas se conoce como regresión en lugar de clasificación.

Para estos problemas de clasificación hay dos tipos de instancias: etiquetadas y no etiquetadas. Una instancia está etiquetada cuando se conoce el valor de la variable clase. Según este tipo de instancias, se pueden definir los siguientes problemas de clasificación:

- *Clasificación supervisada*: problemas de clasificación donde $\mathcal{D}$ sólo contiene instancias etiquetadas con ejemplos de todas las clases posibles.
- *Clasificación semi-supervisada*: problemas de clasificación donde $\mathcal{D}$ contiene instancias etiquetadas con ejemplos de todas las clases y un conjunto de instancias no etiquetadas.
- *Clasificación no supervisada*: problemas de clasificación con todas las instancias no etiquetadas.
- *Clasificación Positive Unlabelled*: problemas de clasificación donde la clase toma sólo dos valores (positivo y negativo) y $\mathcal{D}$ contiene únicamente instancias positivas y sin etiquetar.

En este trabajo se han utilizado soluciones para el problema de clasificación supervisada y clasificación PU.

2.2. Métodos de clasificación supervisada

A continuación se exponen los métodos de clasificación supervisada empleados en el estudio.

2.2.1. Clasificador naive Bayes

El modelo naive Bayes [17, 18], también conocido como modelo ingenuo de Bayes, es el clasificador más simple de redes Bayesianas, ya que asume la independencia condicional de todas las variables predictoras dada la variable clase (Figura 2.1).

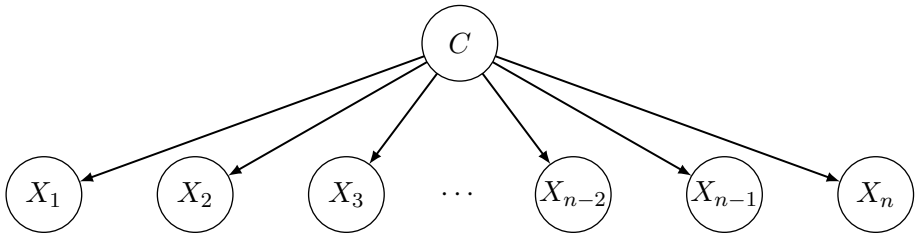

Figura 2.1: Estructura del modelo naive Bayes. Todas las variables predictoras X_i son condicionalmente independientes dada la variable clase C.

Dada su simplicidad, el proceso de aprendizaje es muy rápido incluso para problemas de grandes dimensiones. Esto ocurre porque la estructura

del modelo es fija y no hay necesidad de hacer un paso de aprendizaje estructural; y, porque el número de parámetros que es necesario estimar se reduce debido a que las variables predictoras dependen sólo de la variable clase. Además, cuanto más simple es el modelo, menor es el número de parámetros que hay que estimar y, por lo tanto, más sólidas son las estimaciones de los datos. Probablemente, ésta sea una de las razones por las cuales, independientemente de la fuerte suposición de independencia, los modelos naive Bayes son competitivos con respecto a otros clasificadores más sofisticados.

Los parámetros que hay que estimar a partir de los datos son las probabilidades condicionales, $P(x_{ij}|c)$, y las probabilidades a priori de la clase, $P(c)$. Estas probabilidades se calcularán con los estimadores de máxima verosimilitud. Para evitar la presencia de parámetros nulos que puedan generar problemas cuando se utiliza el clasificador para obtener la probabilidad posterior de la clase, se utiliza la corrección de Laplace [19]:

$$P(x_{ij}|c) = \frac{1 + N_{ijc}}{r_i + N_c} \tag{2.1}$$

$$P(c) = \frac{1 + N_c}{r_c + N} \tag{2.2}$$

donde N_{ijc} es el número de instancias de la clase c con $X_i = x_{ij}$ y N_c es el número de instancias etiquetadas como c. Dado el supuesto de la independencia condicional y la regla de Bayes, por el clasificador naive Bayes se tiene que:

$$P_{nB}(c|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|c)P(c)}{P(\boldsymbol{x})} = \frac{\prod_{i=1}^{n} P(x_i|c)P(c)}{P(\boldsymbol{x})} \approx \prod_{i=1}^{n} P(x_i|c)P(c) \tag{2.3}$$

2.2.2. Evaluación de los clasificadores

Medidas de rendimiento del clasificador

Existen diversas métricas para evaluar la calidad de las funciones de clasificación. Dado que el problema que se está estudiando es de clasificación binaria[3], se puede definir un amplio conjunto de medidas de rendimiento basadas en la matriz de confusión (Tabla 2.1) [20]. Algunas de las métricas más usadas se muestran en la Tabla 2.2 [20].

El objetivo del problema es construir una competitiva función de clasificación probabilística, φ, que, para una nueva instancia $\boldsymbol{x}$, estime la probabilidad de cada una de las clases dados los valores de los atributos de la instancia $P_{nB}(c|\boldsymbol{x})$ y asignarle la clase con la mayor probabilidad posterior.

$$\varphi(\boldsymbol{x}) = argmax_c(P_{nB}(c|\boldsymbol{x}))$$

[3]En un problema de clasificación binaria sólo se tienen dos clases: positiva y negativa, representadas por 1 y 0, respectivamente

		Predicción	
		P	**N**
Valor real	**P**	True Positive	False Negative
	N	False Positive	True Negative

Tabla 2.1: Matriz de confusión en problemas de clasificación binarios [20].

Medida	Cálculo	Interpretación
Sensibilidad, S_n Recall, r Tasa Positiva Real, TPR	$S_n = \frac{TP}{TP+FN}$	Proporción de casos positivos reales que están correctamente clasificados
Especificidad, S_p	$S_p = \frac{TN}{TN+FP}$	Proporción de casos negativos reales que están correctamente clasificados
Tasa de Falsos Positivos, TFP	$FPR = \frac{FP}{TN+FP}$ $= 1 - S_p$	Proporción de casos negativos reales que se han clasificado como positivos
Precisión, p_r	$p_r = \frac{TP}{TP+FP}$	Probabilidad de que una instancia clasificada como positiva sea realmente positiva

Tabla 2.2: Medidas de rendimiento calculadas a partir de la matriz de confusión [20].

Esto significa que, en problemas de clasificación binarios, por defecto se infiere que la instancia $\boldsymbol{x}$ es positiva cuando $P(c = 1|\boldsymbol{x}) > 0.5$. Sin embargo, se puede establecer cualquier número entre 0 y 1 como umbral, dando lugar a diferentes funciones de clasificación. La curva Característica Operativa del Receptor (del inglés, *Receiver Operating Characteristic*, ROC) es una representación gráfica de la sensibilidad frente a la especificidad que permite evaluar el modelo considerando todos los umbrales posibles para la probabilidad posterior. La Figura 2.2 muestra un ejemplo de curva ROC. La curva ROC permite comparar gráficamente diferentes clasificadores simplemente observando la forma de sus curvas. El área bajo la curva ROC (del inglés, *Area Under the Curve*, AUC) resume la curva y da una idea de la bondad del modelo. Cuanto más cerca esté dicho valor de 1, mejor será el modelo obtenido.

En el contexto de recuperación de información, son especialmente intere-

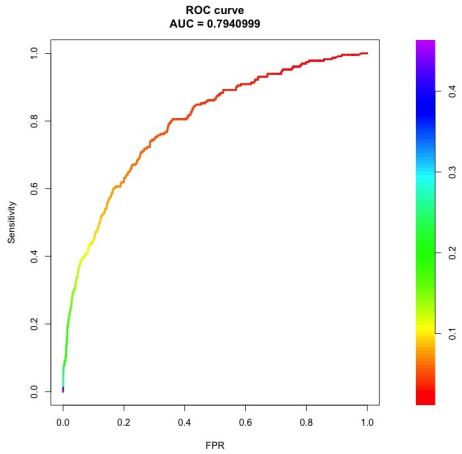

Figura 2.2: Ejemplo de curva ROC.

santes los ejemplos positivos. El objetivo de una función de clasificación es recuperar los casos positivos del conjunto de instancias no etiquetadas. Los dos aspectos de esta recuperación, la cantidad y la calidad, se miden mediante el *recall* y la *precisión*, respectivamente. El recall da una idea de cuántos de los casos positivos reales se han identificado como positivos y la precisión dice cuántas de las instancias predichas como positivas son realmente positivas [20, 21].

Cualquiera de estas medidas por sí sola no es suficiente para evaluar un clasificador, ya que una función que siempre clasifica como *positivo* tendría un recall de 1, aunque el clasificador sea completamente inútil. Lo mismo ocurre con una función que predice un caso único positivo real como positivo, que obtendría una precisión de 1.

La *medida F* es una combinación de estas dos medidas y, por lo tanto, da una idea global del rendimiento de la función de clasificación en la tarea de recuperación de información [20, 21]. Esta medida es la media armónica de la precisión y el recall, y su definición más general, $F_\alpha(r, pr)$, es:

$$\frac{1}{F_\alpha(r,pr)} = \frac{1}{\alpha+1}\left(\frac{\alpha}{r} + \frac{1}{p_r}\right), \quad \alpha \in (0, +\infty) \tag{2.4}$$

El factor de ponderación α permite poner más énfasis en la cantidad o la calidad de la recuperación. La definición más utilizada de medida F toma el factor de ponderación $\alpha = 1$:

$$\frac{1}{F(r,pr)} = \frac{1}{2}\left(\frac{1}{r} + \frac{1}{p_r}\right) \quad \Longrightarrow \quad F(r,pr) = \frac{2rp_r}{r+p_r} \tag{2.5}$$

Estimación de las medidas de rendimiento

Estas medidas se estimarán a partir de los datos utilizando un esquema de validación cruzada repetida de r iteraciones por k hojas.

La validación cruzada de k hojas (del inglés, k-*cross validation*, k-cv) [22] (ver Figura 2.3), es un enfoque que permite estimar la bondad del clasificador sin el problema del sobreajuste de los datos y evitando la reducción del conjunto de entrenamiento. Consiste en dividir aleatoriamente el conjunto de datos en k subconjuntos. Luego, se entrena un modelo con todos los subconjuntos excepto uno, y la medida se calcula en ese subconjunto que se ha omitido. Este procedimiento se repite k veces, dejando fuera en cada paso uno de los subconjuntos, y la medida final se estima como el promedio de los k resultados. El sesgo y la varianza de la estimación dependen de la k utilizada. Los valores más comúnmente usados para k son 5 y 10. En los experimentos realizados se trabaja con $k = 10$.

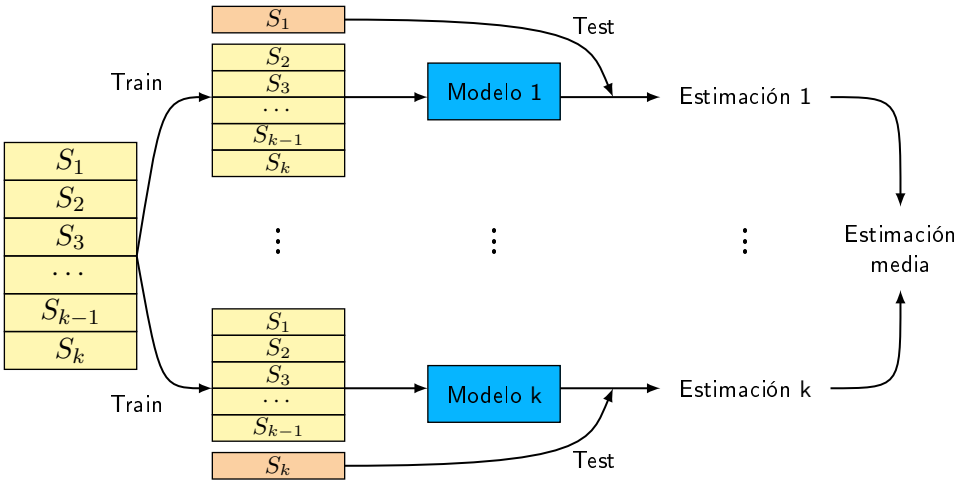

Figura 2.3: Esquema de la k-cv de la estimación de medidas

Sin embargo, la estimación obtenida mediante la k-cv depende de la partición del conjunto de datos. Para superar este problema, se ha usado un k-cv repetido, que consiste en repetir el k-cv r veces. La estimación final es el promedio de las estimaciones individuales $r \cdot k$.

2.2.3. Selección del subconjunto de características

En este capítulo, se ha introducido el concepto de clasificación supervisada, donde las funciones de clasificación predicen la clase de una nueva instancia en función de los valores de sus atributos.

Aunque puede parecer que cuantas más características haya para describir la instancia, mejor se podrá predecir su clase, esa relación no es directamente proporcional. De hecho, no todas las características son igualmente

útiles para el propósito de la clasificación.

Puede haber características que no tienen nada que ver con la clase. Por ejemplo, si se trata de predecir el embarazo de un ciclo de ART, poco importa el nombre de la paciente. Por lo tanto, se pueden dividir las variables de predicción en relevantes (cuando tienen algún poder predictivo) e irrelevantes (cuando no hay relación entre ellas y la clase). También se pueden encontrar dos o más variables que aporten una información similar. Por ejemplo, la fecha de nacimiento y la edad son dos variables directamente correlacionadas. Cuando dos variables están altamente correlacionadas, se dice que son redundantes.

Las variables no informativas y redundantes, además de conducir a modelos demasiado complejos y con un mayor tiempo computacional requerido para aprender el clasificador, pueden ser dañinas para algunos algoritmos de inducción de modelos. Por lo tanto, producir un reducido conjunto de variables relevantes y no redundantes es un paso crucial en muchas aplicaciones de aprendizaje automático [11].

Para reducir la dimensionalidad del problema se usará el paradigma de selección de subconjuntos de características (del inglés, *Feature Subset Selection*, FSS) [23, 21, 24]. En particular, se utilizará un método de filtro multivariante conocido como la selección de características basadas en la correlación (del inglés, *Correlation-based Feature Selection*, CFS) [23].

El algoritmo CFS es un algoritmo para filtrar variables irrelevantes y redundantes. Este método busca el mejor subconjunto de características guiado por una métrica que mide tanto la correlación entre cada variable con la clase y la correlación entre las variables seleccionadas. Esta métrica multivariada mide la bondad del subconjunto candidato de variables de predicción en términos de la relevancia y la no redundancia de las mismas. En un subconjunto, la relevancia se mide como la correlación entre cada variable del conjunto y la variable clase, mientras que la redundancia se mide como la correlación entre cada par de variables de predicción.

El objetivo es obtener un subconjunto de variables relevantes (características fuertemente correlacionadas con la clase) sin redundancias (características con una pequeña correlación entre ellas).

Una de las ventajas de este método es que es independiente del paradigma de clasificación [23]. Por lo tanto, una vez que se tiene un subconjunto de características, éste puede usarse en el entrenamiento de cualquier tipo de modelo.

La métrica propuesta por Hall y Smith [23] para evaluar un conjunto dado S se define de la siguiente manera:

$$G_S = \frac{k\overline{u_{ci}}}{\sqrt{k + k(k-1)\overline{u_{ii'}}}} \tag{2.6}$$

donde k es el número de variables en S, $\overline{u_{ci}}$ es la correlación promedio entre las características en S y la clase, y $\overline{u_{ii'}}$ es la correlación promedio entre las características incluidas en S.

La correlación entre cada par de variables se mide en términos del coeficiente de incertidumbre $U(X_k|X_i)$ [25], que es una medida propuesta en el marco de la teoría de la información y se basa en la información mutua $I(X_k; X_i)$ y la entropía $H(X_k)$. Cuando X_i y X_k son variables discretas, se define como:

$$U(X_k|X_i) = \frac{I(X_k; X_i)}{H(X_k)} = \frac{H(X_k) - H(X_k|X_i)}{H(X_k)} \tag{2.7}$$

$$H(X_k) = -\sum_{l=1}^{r_k} P(x_{kl})\ log(P(x_{kl})) \tag{2.8}$$

$$H(X_k|X_i) = -\sum_{j=1}^{r_i} P(x_{ij}) \sum_{l=1}^{r_k} P(x_{kl}|x_{ij})\ log(P(x_{kl}|x_{ij})) \tag{2.9}$$

Mirando la métrica G_S se observa que cuanto mayor es la correlación media de las características con la clase (i.e., la relevancia promedio de las características) mayor es la métrica, y cuanto mayor es la correlación entre las variables (i.e., mayor es la redundancia) menor es la métrica. Por lo tanto, dado un conjunto de variables, la adición de variables redundantes disminuirá el *score*, mientras que la adición de características relevantes la aumentará.

Dado un conjunto de variables, la búsqueda del subconjunto de características que maximiza la métrica G_S se puede concebir como un problema de optimización. La búsqueda en el espacio del subconjunto de características es un problema NP-difícil[4]. Este problema se supera usando heurísticas de búsqueda que pueden encontrar buenas y subóptimas soluciones en un tiempo computacional razonable. El pseudocódigo de la selección hacia adelante aplicada a la selección de características se encuentra en la Figura 2.4.

Pseudocódigo de la selección avariciosa de ascenso ascendente en el algoritmo CFS

2.3. Métodos de clasificación PU

A continuación se exponen los métodos de clasificación PU empleados en el estudio.

2.3.1. El algoritmo EM

El algoritmo EM (del inglés, *Expectation–Maximization*) [14, 26] es un popular algoritmo iterativo para la estimación de máxima verosimilitud en

[4]Es decir, a medida que aumenta el número de características, la búsqueda exhaustiva se hace inviable computacionalmente

Inicialización

Sea $S_0 = X_i$ la solución inicial, donde X_i es la característica con la mayor correlación con la variable clase (medida en términos del coeficiente de incertidumbre)

D = Número de variables predictoras de $\mathcal{D}$

$i = 0$

Proceso

Mientras $i \leq D$

Se obtienen todas las soluciones candidatas agregando a S_i cada una de las variables que no están en S_i (una característica en cada solución candidata)

Para cada solución candidata, L_j, se calcula G_{L_j}

Si $argmax_{L_j} G_{L_j} \geq G_{S_i}$

$S_{i+1} = argmax_{L_j} G_{L_j}$

$i = i + 1$

Si no

Parar

Salida: Se devuelve S_i como el mejor subconjunto

Figura 2.4: Pseudocódigo de la selección de variables ascendente en el algoritmo CFS (*greedy hill climbing forward selection*) [21].

problemas con datos incompletos. A menudo se usa para completar los valores perdidos en los datos usando los valores existentes para calcular el valor esperado para cada valor desconocido. El algoritmo EM consta de dos pasos, el paso de Expectativa (E) y el paso de Maximización (M). El paso E básicamente completa los datos que faltan. Los parámetros se estiman en el paso M después de completar o reconstruir los datos faltantes. Esto lleva a la siguiente iteración del algoritmo. Para el clasificador naive Bayes, los pasos utilizados por EM son idénticos a los realizados para construir el clasificador (Ecuaciones 2.1 y 2.2 para el paso E y la Ecuación 2.3 para el paso M). Hay que tener en cuenta que en este algoritmo las instancias tienen una etiqueta de clase probabilística que toma valores en $[0, 1]$, en lugar de la etiqueta de clase $\{0, 1\}$.

Esta capacidad del EM para trabajar con datos perdidos es exactamente lo que se busca para dar solución al problema PU. Dado un conjunto $\mathcal{P}$ de instancias positivas con valor de clase $c = 1$, y otro conjunto $\mathcal{M}$[5] de instancias mixtas sin etiqueta de clase, se busca obtener la clase de cada

[5]El conjunto $\mathcal{M}$ está formado por instancias $\boldsymbol{x}_M$ sin etiqueta de clase asignada, que pueden ser positivas o negativas.

instancia del conjunto mixto. Para facilitar la lectura, en esta sección las instancias del conjunto $\mathcal{P}$ se denotarán por $\boldsymbol{x}_P$, y las del conjunto $\mathcal{M}$ por $\boldsymbol{x}_M$. EM puede ayudar a asignar una etiqueta de clase probabilística a cada ejemplo en el conjunto mixto, es decir, $P(c = 1|\boldsymbol{x}_M)$ y $P(c = 0|\boldsymbol{x}_M)$. Después de varias iteraciones, todas las probabilidades convergerán. Sin embargo, una buena inicialización es importante para encontrar una buena función de verosimilitud. Como sólo se tienen instancias positivas, el principal problema consiste en identificar algunos ejemplos de clase negativa. A continuación, se propone una técnica para identificar los ejemplos negativos más probables, que se utilizarán en la inicialización del EM. Este procedimiento, conocido como *spy-EM* [14], consta de dos pasos principales: I. Reinicialización; y, II. Construcción y selección del clasificador final.

Paso 1: Reinicialización

Aplicación del algoritmo EM Inicialmente, se asigna a cada instancia positiva $\boldsymbol{x}_P \in \mathcal{P}$ la etiqueta de clase $c = 1$ y a cada instancia $\boldsymbol{x}_M$ del conjunto mixto $\mathcal{M}$ la etiqueta de clase $c = 0$; es decir,

$$P(c = 1|\boldsymbol{x}_P) = 1 \text{ y } P(c = 0|\boldsymbol{x}_P) = 0$$

$$P(c = 0|\boldsymbol{x}_M) = 1 \text{ y } P(c = 1|\boldsymbol{x}_M) = 0$$

Con esta etiqueta inicial, se construye un clasificador naive Bayes (NB-C) que se utilizará para clasificar las instancias del conjunto $\mathcal{M}$. El clasificador NB-C se usa para calcular la probabilidad posterior $P(c = 1|\boldsymbol{x}_M)$ de cada instancia en el conjunto mixto (usando la Ecuación 2.3), que se asigna a $\boldsymbol{x}_M$ como su nuevo valor de clase probabilística. La probabilidad de clase para cada ejemplo positivo $\boldsymbol{x}_P$ sigue siendo la misma durante todo el proceso.

Después de asignar un nuevo valor de clase probabilística a las instancias del conjunto mixto, se construye un nuevo clasificador NB-C basado en los nuevos valores $P(c = 1|\boldsymbol{x}_M)$ y $P(c = 1|\boldsymbol{x}_P) = 1$ de los conjuntos $\mathcal{M}$ y $\mathcal{P}$, respectivamente. Empieza la siguiente iteración. Este proceso iterativo continúa hasta que EM converge. El algoritmo completo, llamado EM inicial [14] (del inglés, *Initial EM*, I-EM), se da en la Figura 2.5. Hay que tener en cuenta que durante el proceso de asignación de etiquetas de clase probabilísticas a cada ejemplo $\boldsymbol{x}_M \in \mathcal{M}$, también se puede calcular $P(x_{ij}|c)$ y $P(c)$ (Ecuaciones 2.1 y 2.2), que son suficientes para construir un nuevo NB-C ya que la información calculada en el proceso inicial para el conjunto positivo $\mathcal{P}$ continua siendo la misma.

La etiqueta de clase probabilística final para cada instancia $\boldsymbol{x}_M \in \mathcal{M}$ se usa para clasificar las instancias del conjunto e identificar los ejemplos negativos.

Parámetros de entrada:

$\mathcal{M}$ = Conjunto de instancias mixtas

$\mathcal{P}$ = Conjunto de instancias positivas

Proceso:

Construir un clasificador naive Bayes inicial NB-C usando $\mathcal{P}$ como conjunto positivo y $\mathcal{M}$ como negativo.

Recorrer mientras cambian los parámetros del clasificador

Para cada ejemplo $\boldsymbol{x}_M \in \mathcal{M}$

Calcular $P(c = 1|\boldsymbol{x}_M)$ usando el modelo NB-C actual.
$\Longrightarrow P(c = 0|\boldsymbol{x}_M) = 1 - P(c = 1|\boldsymbol{x}_M)$.
Actualizar $P(x_{ij}|c = 1)$ y $P(c = 1)$ dada la clase probabilística asignada a $\boldsymbol{x}_M$ ($P(c = 1|\boldsymbol{x}_M)$) y P. En el proceso se está construyendo un nuevo NB-C.

Salida: Se devuelve el clasificador final NB-C.

Figura 2.5: El algoritmo I-EM con clasificador naive Bayes.

Introducción de instancias espía en el conjunto mixto Dado que en un principio sólo se tienen ejemplos positivos, la inicialización de la que se parte está fuertemente sesgada hacia instancias positivas. Para resolver ese problema, se necesita una inicialización que equilibre tanto los ejemplos positivos como negativos. Sin embargo, al no saber cuáles son negativos, se deben identificar dentro del conjunto mixto algunas instancias que resulten negativas muy probablemente.

Para obtener información fiable para la identificación de los ejemplos negativos se introducen ejemplos del conjunto positivo $\mathcal{P}$ al conjunto mixto $\mathcal{M}$. Estos ejemplos actuarán como espías del conjunto $\mathcal{P}$. En este enfoque se selecciona al azar un $s\,\%$ de las instancias de $\mathcal{P}$ (en los experimentos realizados, se usa el $10\,\%$), que formarán el conjunto $\mathcal{S}$ de espías, y se agregan al conjunto mixto $\mathcal{M}$. Los espías se comportan de forma idéntica a los ejemplos positivos desconocidos en $\mathcal{M}$ y, por lo tanto, permiten inferir de manera fiable el comportamiento de los ejemplos positivos desconocidos.

Se vuelve a lanzar el algoritmo I-EM para el conjunto $\mathcal{P}$ y el $\mathcal{MS}$ (el conjunto mixto con los espías). Una vez que el algoritmo EM se completa, las etiquetas probabilísticas de los ejemplos de $\mathcal{S}$ se utilizan para decidir qué ejemplos tienen más probabilidades de ser negativos. Para separar los ejemplos de $\mathcal{M}$ en negativos fiables ($\mathcal{RN}$) y sin etiquetar ($\mathcal{U}$) se emplea un umbral t. Los ejemplos en $\mathcal{M}$ con probabilidades más bajas que t son los ejemplos negativos más probables y el resto de ejemplos con probabilidades más altas que t se convierten en ejemplos no etiquetados. El algoritmo detallado para identificar los conjuntos $\mathcal{RN}$ y $\mathcal{U}$ se muestra en la Figura 2.6.

Parámetros de entrada:

$\mathcal{M}$ = Conjunto mixto de instancias sin etiquetar

$\mathcal{P}$ = Conjunto de instancias positivas

Pasos a seguir:

1. $\mathcal{N} = \mathcal{U} = \emptyset$
2. $\mathcal{S} = muestra(\mathcal{P}, s\,\%)$
3. $\mathcal{MS} = \mathcal{M} \cup \mathcal{S}$
4. $\mathcal{P} = \mathcal{P} \setminus \mathcal{S}$
5. A todas las instancias $\boldsymbol{x}_P \in \mathcal{P}$ se les asigna la etiqueta de clase $c = 1$
6. A todas las instancias $\boldsymbol{x}_M \in \mathcal{MS}$ se les asigna la etiqueta de clase $c = 0$
7. **Ejecutar** I-EM($\mathcal{MS}, \mathcal{P}$)
8. Clasificar cada instancia de $\mathcal{MS}$
9. Determinar el umbral de probabilidad t usando $\mathcal{S}$
10. Separar el conjunto $\mathcal{M}$ en los conjuntos $\mathcal{RN}$ y $\mathcal{U}$:
11. **Para cada instancia** $\boldsymbol{x}_M \in \mathcal{M}$
12. **Si** $P(c = 1|\boldsymbol{x}_M) < t$
13. $\mathcal{N} = \mathcal{N} \cup \{\boldsymbol{x}_M\}$
14. **Si no**
15. $\mathcal{U} = \mathcal{U} \cup \{\boldsymbol{x}_M\}$

Salida: Se devuelve el conjunto de los negativos fiables $\mathcal{RN}$ y el conjunto de no etiquetados $\mathcal{U}$.

Figura 2.6: Pseudocódigo del proceso de identificación de ejemplos negativos fiables.

Ahora se discute cómo determinar el umbral t [14]. Sea $\mathcal{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_k\}$ el conjunto de espías, y sea $P(c_1|\boldsymbol{s}_i)$ la etiqueta probabilística asignada a cada $\boldsymbol{s}_i$. Intuitivamente, se puede usar la probabilidad mínima de $\mathcal{S}$ como valor de umbral t, es decir, $t = min\{P(c_1|\boldsymbol{s}_1), P(c_1|\boldsymbol{s}_2), \ldots, P(c_1|\boldsymbol{s}_k)\}$, lo que significa que queremos recuperar todos los ejemplos espía. En un dominio sin ruido, usar la probabilidad mínima es aceptable. Sin embargo, la mayoría de las colecciones de ejemplos de la vida real tienen valores atípicos y ruido, por lo que usar la probabilidad mínima no es fiable. No obstante, se desconoce el nivel de ruido de los datos. Éste se puede estimar probando distintos niveles de ruido y seleccionando el mejor. Primero se ordenan los ejemplos en $\mathcal{S}$ de acuerdo con su $P(c_1|s_i)$ y se usa un nivel de ruido l para decidir el umbral t. Se selecciona t tal que el $l\,\%$ de los ejemplos tienen una probabilidad menor

que t. En los experimentos realizados, se han probado varios niveles de ruido ($l = 5$, 10, 15, 20 y 25). Los resultados no muestran grandes diferencias para los niveles elegidos (como se verá al discutir el Paso 2).

En resumen, el objetivo de este primer paso del algoritmo propuesto es lograr los resultados que se muestran en la Figura 2.7. La parte izquierda muestra la situación inicial, donde al conjunto mixto, formado por ejemplos positivos y negativos sin conocer su identidad, se le agregan los espías de $\mathcal{P}$. La parte derecha de la figura, muestra el resultado que la técnica logra con la ayuda de espías. Se observa que la mayoría de los ejemplos positivos en el conjunto mixto se colocan en el conjunto no etiquetado y una parte importante de los ejemplos negativos se colocan en el conjunto $\mathcal{RN}$ de negativos más fiables. La pureza de $\mathcal{RN}$ es mucho más alta que la del conjunto mixto.

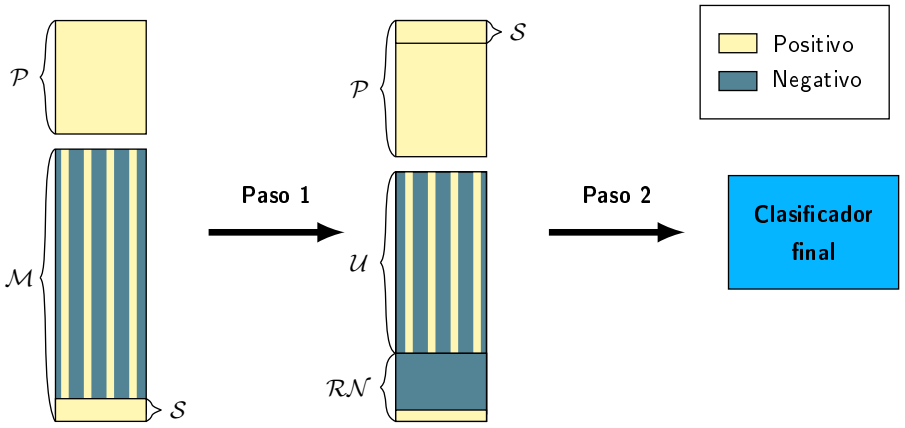

Figura 2.7: Ilustración de la estrategia de dos pasos spy-EM.

Paso 2: Construcción del clasificador final

En este paso se construye el clasificador final. El algoritmo EM se emplea de nuevo, con los conjuntos $\mathcal{P}$, $\mathcal{RN}$ y $\mathcal{U}$. Este paso se lleva a cabo de la siguiente manera:

1. Se devuelven todos los ejemplos espías al conjunto positivo $\mathcal{P}$
2. Se asigna a todos los ejemplos $\boldsymbol{x}_P$ en el conjunto positivo $\mathcal{P}$ la etiqueta de clase fija $c = 1$, $P(c = 1|\boldsymbol{x}_P) = 1$, que no cambiará en cada iteración del EM.
3. Se asigna a todos los ejemplos $\boldsymbol{x}_{RN}$ en el conjunto de negativos fiables $\mathcal{RN}$ la etiqueta de clase inicial $c = 0$, $P(c = 0|\boldsymbol{x}_{RN}) = 1$, que cambia con cada iteración del EM.
4. Cada ejemplo $\boldsymbol{x}_U$ en el conjunto sin etiquetar $\mathcal{U}$ no tiene asignada ninguna etiqueta inicialmente. Al final de la primera iteración del EM, se le

asignará una etiqueta probabilística, $P(c = 1|\boldsymbol{x}_U)$. En iteraciones posteriores, el conjunto $\mathcal{U}$ participará en EM con sus clases probabilísticas recientemente asignadas.

5. Se ejecuta el algoritmo EM utilizando los conjuntos $\mathcal{P}$, $\mathcal{RN}$ y $\mathcal{U}$ hasta que converja y se obtenga el clasificador final.

Como se ha señalado anteriormente, el valor de l, el porcentaje utilizado para seleccionar los ejemplos negativos $\mathcal{RN}$ más fiables, no es crítico siempre que se encuentre en un rango razonable desde $5\,\% - 25\,\%$. Esto es debido a que en el Paso 2 del algoritmo, las probabilidades de los conjuntos $\mathcal{U}$ y $\mathcal{RN}$ pueden variar. Si hay demasiados ejemplos positivos en $\mathcal{RN}$, el algoritmo EM corregirá lentamente la situación, es decir, los mueve al lado positivo. En los experimentos realizados, se usa $l = 15\,\%$, lo cual funcionó muy bien. También se ha experimentado con $5\,\%$, $10\,\%$, $20\,\%$ y $25\,\%$ obteniendo resultados similares.

2.3.2. Clasificador naive Bayes positivo

El algoritmo naive Bayes positivo (del inglés, *Positive Naive Bayes*, PNB) [27] es una adaptación del modelo naive Bayes de clasificación supervisada (Sección 2.2.1) capaz de aprender modelos naive Bayes a partir de un conjunto de datos de entrenamiento $\mathcal{D}$ de instancias positivas y no etiquetadas ($\mathcal{D} = \mathcal{D}_P \cup \mathcal{D}_U$). Como ya se ha mencionado, un modelo naive Bayes asume que todas las variables de predicción son condicionalmente independientes dada la variable de clase. Dada esta suposición, los parámetros que se necesitan estimar son las probabilidades condicionales $P(x_{ij}|c)$ y las probabilidades a priori de la variable de clase $P(c)$. Cuando la variable clase C toma sólo dos valores las probabilidades condicionales se pueden dividir en dos grupos, $P(x_{ij}|c = 1)$ y $P(x_{ij}|c = 0)$[6]. La probabilidad a priori de la clase positiva ($P(c = 1)$, que a partir de ahora se representará como p para simplificar) se estima a partir de todo el conjunto de datos $\mathcal{D}$.

En el marco de aprendizaje PU no se dispone de ejemplos negativos y, por lo tanto, los parámetros relacionados con la clase negativa no pueden estimarse desde $\mathcal{D}$. Dado que:

$$P(x_{ij}) = P(x_{ij}|c = 1)p + P(x_{ij}|c = 0)(1 - p) \tag{2.10}$$

se tiene que $P(x_{ij}|c = 0)$ se puede estimar como:

$$P(x_{ij}|c = 0) = \frac{P(x_{ij}) - P(x_{ij}|c = 1)p}{1 - p} \tag{2.11}$$

[6]En los problemas de aprendizaje PU sólo hay dos clases, a partir de ahora se asume que la clase C es una variable aleatoria binaria, donde 0 y 1 representarán la clase negativa y la positiva, respectivamente.

donde $P(x_{ij})$ puede estimarse a partir de las instancias no etiquetadas ya que no depende de la clase. El parámetro p no puede estimarse sin ejemplos negativos y, por lo tanto, debe ser introducido como parámetro por el usuario. De hecho, éste es el principal inconveniente del PNB, ya que si se conociera el valor de esta probabilidad, la estimación de los parámetros sería la misma que en el algoritmo supervisado.

$P(x_{ij})$ se puede estimar a partir de las instancias no etiquetadas como:

$$P(x_{ij}) = \frac{N_{ij\mathcal{D}_U}}{N_{\mathcal{D}_U}} \tag{2.12}$$

donde $N_{\mathcal{D}_U}$ es el número de instancias no etiquetadas y $N_{ij\mathcal{D}_U}$ el número de instancias no etiquetadas donde $X_i = x_{ij}$. Reemplazando esta probabilidad en la Ecuación (2.11), se tiene:

$$P(x_{ij}|c=0) = \frac{N_{ij\mathcal{D}_U} - P(x_{ij}|c=1)pN_{\mathcal{D}_U}}{(1-p)N_{\mathcal{D}_U}} \tag{2.13}$$

El problema con este estimador es que puede ser negativo porque, aunque $P(x_{ij}) \geq P(x_{ij}|c=1)p$, el sesgo en la estimación de estas probabilidades puede conducir a situaciones donde la estimación de $P(x_{ij}|c=1)p$ es mayor que la estimación de $P(x_{ij})$, arrojando una estimación negativa de $P(x_{ij}|c=0)$. Para resolver este problema, Denis et al. [27] proponen reemplazar las estimaciones negativas por cero y luego normalizar las probabilidades de modo que para cada variable predictora X_i, $\sum_{j=1}^{r_i} P(x_{ij}|c=0) = 1$. Podemos definir el factor de normalización Z_i para la variable de predicción X_i como:

$$Z_i = \sum_{j=1}^{r_i} max\left(0; \frac{N_{ij\mathcal{D}_U} - P(x_{ij}|c=1)pN_{\mathcal{D}_U}}{(1-p)N_{\mathcal{D}_U}}\right) \tag{2.14}$$

Después de la normalización, el estimador de las probabilidades relacionadas con la clase negativa sería:

$$P(x_{ij}|c=0) = \frac{max(0; R_i(j))\frac{1}{Z_i}}{(1-p)N_{\mathcal{D}_U}} \tag{2.15}$$

$$R_i(j) = N_{ij\mathcal{D}_U} - P(x_{ij}|c=1)pN_{\mathcal{D}_U} \tag{2.16}$$

Finalmente, introduciendo la corrección de Laplace [19], se tiene que:

$$P(x_{ij}|c=0) = \frac{1 + max(0; R_i(j))\frac{1}{Z_i}}{r_i + (1-p)N_{\mathcal{D}_U}} \tag{2.17}$$

verificando que $\sum_{j=1}^{r_i} P(x_{ij}|c=0) = 1$.

Para resumir, el algoritmo PNB estima los parámetros de un modelo naive Bayes de la siguiente manera:

- $P(x_{ij}|c=1)$ se estima a partir de los ejemplos positivos.
- $P(x_{ij}|c=0)$ se estima utilizando la Ecuación 2.17.
- p es un parámetro del algoritmo.

2.3.3. Evaluación de los clasificadores

Además del problema de inducción del clasificador, la ausencia de casos negativos obliga a establecer la probabilidad a priori de la clase positiva p, ya que no se puede estimar a partir de los datos. A continuación se definen unas métricas para poder estimar la medida F real con ejemplos positivos y no etiquetados y una forma de evaluar o, al menos, comparar clasificadores en el contexto de aprendizaje PU.

Medidas de rendimiento en el contexto de aprendizaje PU

Uno de los principales problemas en el contexto de aprendizaje positivo no etiquetado es la evaluación de los clasificadores. La ausencia de ejemplos negativos hace que sea imposible estimar casi ninguna de las medidas de rendimiento clásicas (Tabla 2.2). De hecho, la única que es posible estimar es el recall, ya que sólo se ocupa de los ejemplos positivos. Todas las medidas de rendimiento, como la precisión o la medida F, se estiman utilizando la información contenida en la matriz de confusión (Tabla 2.1).

Desde un punto de vista probabilístico, el recall es la probabilidad de que un caso positivo sea clasificado correctamente como positivo por la función de clasificación y la precisión es la probabilidad de que una instancia que ha sido clasificada como positiva sea realmente positiva. Por lo tanto, se tiene que:

$$r = P(\varphi(\boldsymbol{x}) = 1|c = 1) \tag{2.18}$$

$$p_r = P(c = 1|\varphi(\boldsymbol{x}) = 1) \tag{2.19}$$

Teniendo en cuenta la regla de Bayes, se tiene:

$$p_r = P(c = 1|\varphi(\boldsymbol{x}) = 1) = \frac{P(\varphi(\boldsymbol{x}) = 1|c = 1)p}{P(\varphi(\boldsymbol{x}) = 1)} \tag{2.20}$$

donde $P(\varphi(\boldsymbol{x}) = 1)$ es la probabilidad de que el clasificador clasifique una instancia como positiva. Esta probabilidad se puede obtener clasificando todas las instancias posibles y luego obteniendo la proporción de instancias clasificadas como positivas.

En la Sección 2.2.2 se vio que la medida F se define como:

$$F(r, p_r) = \frac{2rp_r}{r + p_r} \tag{2.21}$$

Combinando las Ecuaciones (2.20) y (2.21), se tiene que:

$$F = \frac{2r\frac{rp}{P(\varphi(\boldsymbol{x})=1)}}{r\left(1+\frac{p}{P(\varphi(\boldsymbol{x})=1)}\right)} = \frac{2rp}{P(\varphi(\boldsymbol{x})=1)+p} \tag{2.22}$$

En el contexto PU, el recall se puede estimar con un esquema k-cv como se muestra en la Figura 2.8.

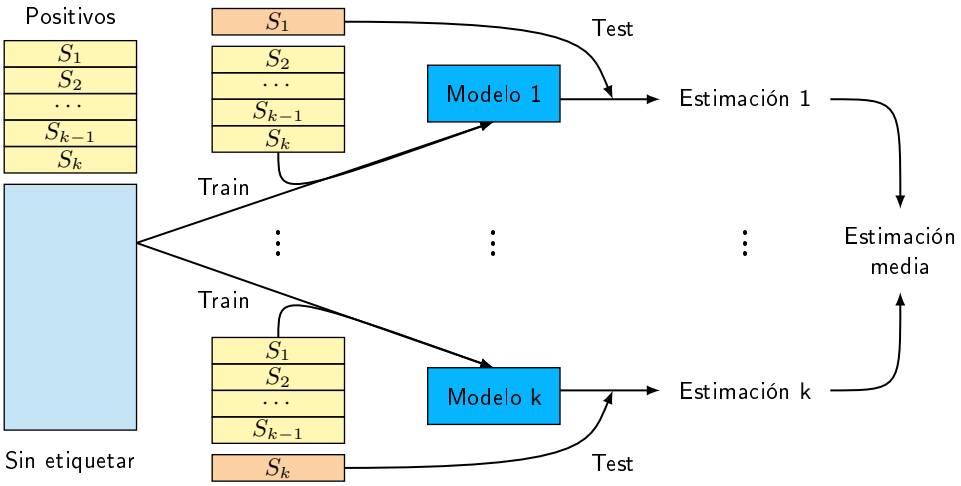

Figura 2.8: Esquema de la estimación k-cv de medidas en el aprendizaje PU.

Con la función de clasificación φ, se puede calcular el valor $P(\varphi(\boldsymbol{x}) = 1)$. El problema es el parámetro p. Si se conociera la probabilidad a priori de la clase positiva, se podría estimar la medida F en problemas de aprendizaje PU.

El problema con el estimador mostrado en la Ecuación 2.22 es que depende del parámetro p. Esta dependencia se debe principalmente a la p en el numerador que hace que la estimación sea proporcional a la probabilidad a priori de la clase positiva. Para resolver parcialmente el problema de la fuerte dependencia del parámetro p, Calvo [21] propone una nueva métrica, la *pseudo-F* (F_{ps}), que es proporcional a la medida F real, sin depender tanto de p:

$$F_{ps} = \frac{r}{P(\varphi(\boldsymbol{x}) = 1) + p} \tag{2.23}$$

Dado que esta métrica es proporcional a la medida F, se comporta igual que ella; y, por lo tanto, se puede usar para buscar el valor p que maximiza la medida F real (a partir de ahora esta p se denominará como la p óptima).

Clasificador naive Bayes positivo wrapper

Como ya se ha visto, la medida pseudo-F se puede usar para identificar el valor de p que maximiza el promedio de la precisión y el recall. Por lo tanto,

esta métrica se usará para definir los clasificadores naive Bayes positivos wrapper (del inglés, *wrapper positive Bayesian network classifiers*, wPBC), que, dado un conjunto de ejemplos positivos y no etiquetados, construyen un clasificador que establece el parámetro p en su valor óptimo por el conjunto de datos y optimiza el recall (en términos de la medida F).

La Figura 2.9 muestra el pseudocódigo de la versión wrapper del PNB (wPNB) utilizando la medida de pseudo-F como métrica de guía [21].

Parámetros de entrada:

I = Número de intervalos

K = Número de rondas

p_{min} = Valor mínimo de p

p_{max} = Valor máximo de p

Proceso:

De 1:K

$p_1 = p_{min}$; $p_I = p_{max}$

$\forall i = 2, \ldots, I-1$, $p_i = p_{i-1} + \frac{p_I - p_1}{I}$

$\forall i = 1, \ldots, I$, se genera un modelo M_i usando el algoritmo PNB con p_i como el parámetro p

$\forall M_i, i = 1, \ldots, I$, estimar el recall r_i y $P(\varphi_{M_i}(\boldsymbol{x}) = 1)$

$\forall M_i, i = 1, \ldots, I$, estimar la métrica pseudo-F:

$F_{ps}^{(i)} = \frac{r_i}{P(\varphi_{M_i}(\boldsymbol{x})=1)+p}$

$opt = argmax_i\{F_{ps}^{(i)}\}$

Si $opt \neq 1$

$p_1 = p_{opt-1}$

Si $opt \neq I$

$p_I = p_{opt+1}$

Salida: Se devuelve el clasificador wPNB M_{opt}.

Figura 2.9: Pseudocódigo del algoritmo wPNB utilizando pseudo-F como métrica de guía [21].

2.3.4. Selección del subconjunto de características en el contexto PU

Dado que los algoritmos existentes de selección de subconjuntos de características (FSS) suelen depender de estimaciones que involucran la variable clase, la ausencia de instancias negativas implica que no se pueden aplicar

directamente en este contexto. Para dar solución a este problema, Calvo [21] introdujo la adaptación del algoritmo CFS al contexto de la clasificación PU.

El algoritmo CFS en ausencia de ejemplos negativos En el marco de clasificación supervisada, se tiene un conjunto de datos de instancias que contienen ejemplos de todas las clases. CFS usa estos ejemplos para obtener, para un subconjunto de características candidatas, la correlación entre cada par de características y entre cada característica y la variable clase.

En el contexto de aprendizaje PU, al tratar de estimar las correlaciones requeridas por el CFS del conjunto de ejemplos positivos y no etiquetados, se observa que la correlación entre las características puede estimarse a partir de los casos no etiquetados (ya que es independiente de la variable clase), pero la correlación entre cada característica y la clase no puede estimarse a partir de los datos sin ejemplos negativos.

El problema surge cuando se trata de obtener $U(X_i|C)$:

$$U(X_i|C) = \frac{I(X_i;C)}{H(X_i)} = \frac{H(X_i) - H(X_i|C)}{H(X_i)} \tag{2.24}$$

En particular, el problema reside en tratar de estimar la entropía condicional $H(X_i|C)$, ya que la entropía de la variable de predicción X_i no depende de la variable clase y, por lo tanto, se puede obtener incluyendo también el conjunto de instancias no etiquetadas.

Este problema puede salvarse descomponiendo el estimador de la entropía condicional en dos términos, uno que contiene las probabilidades relacionadas con la clase positiva y el otro las probabilidades relacionadas con la negativa:

$$\begin{aligned} H(X_i|C) = &- p \sum_{j=1}^{r_i} P(x_{ij}|c=1) log(P(x_{ij}|c=1)) \\ &- (1-p) \sum_{j=1}^{r_i} P(x_{ij}|c=0) log(P(x_{ij}|c=0)) \end{aligned} \tag{2.25}$$

donde $P(x_{ij}|c=1)$ se puede estimar a partir de los ejemplos positivos. Como ocurre en el aprendizaje de los clasificadores de red Bayesianos a partir de ejemplos positivos y no etiquetados, ni p ni $P(x_{ij}|c=0)$ pueden obtenerse directamente de los conjuntos de datos, pero se puede expresar $P(x_{ij}|c=0)$ como:

$$P(x_{ij}|c=0) = \frac{P(x_{ij}) - P(x_{ij}|c=1)p}{1-p} \tag{2.26}$$

Por lo tanto, se puede estimar $P(x_{ij}|c=0)$ basado en p como:

$$\frac{N_{ij\mathcal{D}_U} - P(x_{ij}|c=1)pN_{\mathcal{D}_U}}{(1-p)N_{\mathcal{D}_U}} \tag{2.27}$$

donde $N_{ij\mathcal{D}_U}$ es la cantidad de instancias no etiquetadas con $X_i = x_{ij}$ y $N_{\mathcal{D}_U}$ la cardinalidad del conjunto de ejemplos no etiquetados. El problema con este estimador es que puede ser negativo. Como en el caso de los PBC, las estimaciones negativas se reemplazan por 0, y luego todas las probabilidades se normalizan de modo que $\forall i$, $\sum_{j=1}^{r_j} P(x_{ij}|c=0) = 1$. Después de la normalización y teniendo en cuenta la corrección de Laplace [19], $P(x_{ij}|c=0)$ se puede estimar como:

$$P(x_{ij}|c=0) = \frac{1 + max(0; R_i(j))\frac{1}{Z_i}}{r_i + (1-p)N_{\mathcal{D}_U}} \tag{2.28}$$

$$R_i(j) = N_{ij\mathcal{D}_U} - P(x_{ij}|c=1)pN_{\mathcal{D}_U} \tag{2.29}$$

$$Z_i = \sum_{j=1}^{r_i} max\left(0; \frac{R_i(j)}{(1-p)N_{\mathcal{D}_U}}\right) \tag{2.30}$$

Como p no se puede estimar sólo a partir de ejemplos positivos y no etiquetados, permanece como un parámetro del algoritmo.

Tomando esta ecuación en cuenta, y una vez que se establece el valor de p, se puede estimar la entropía de cualquier variable de predicción condicionada a la variable de clase y, así, obtener el *score* G_S (Ecuación 2.6) para un conjunto dado de características de un conjunto de datos sin ejemplos negativos. El algoritmo CFS donde la métrica G_S se calcula de esta manera se denomina CFS positivo no etiquetado (del inglés, *positive unlabelled CFS*, puCFS).

Capítulo 3

Resultados y análisis

En este capítulo, se mostrarán y analizarán los resultados obtenidos de los clasificadores construidos con los diferentes enfoques y se extraerán conclusiones sobre la influencia de las variables agregadas.

El objetivo de estos experimentos es mostrar la capacidad predictiva de los datos recogidos en los laboratorios para apoyar el proceso de toma de decisiones de los médicos en su práctica diaria. Para ello, se propone una solución basada en técnicas de aprendizaje automático. Teniendo en cuenta toda la información recogida por los médicos, se aprenden clasificadores que podrían ser utilizados para mejorar la tasa de embarazo de las ART. Para construir los modelos se han considerado, en un principio, todas las características recogidas. En otros modelos sólo se utiliza un subconjunto de variables relevantes y no redundantes para el aprendizaje del clasificador. Del mismo modo, se utiliza una aproximación PU que aprovecha los ciclos de destino desconocido para aprender el modelo.

En los experimentos, todas las variables continuas han sido discretizadas usando la misma frecuencia con 3 intervalos. Para obtener el conjunto reducido de variables no redundantes y altamente correlacionadas con la variable clase, se ha aplicado la selección de características basadas en la correlación (del inglés, *Correlation-based Feature Selection*, CFS) [23, 21]. Los métodos implementados para la predicción *embarazable*, que se basan en la estrategia EM iterativa, se detienen cuando la diferencia de media relativa entre los parámetros de dos modelos aprendidos en iteraciones consecutivas es inferior al 0.1 %. Todos los experimentos han sido validados mediante una 10-cv repetida 5 veces. Por razones de claridad, en este capítulo sólo se muestran los resultados más relevantes de cada configuración experimental para cada clasificador NB-C. Las tablas completas de resultados están en el Apéndice A.

Los resultados se van a analizar desde diferentes perspectivas. (1) Según la aproximación: supervisado (embarazo) vs. PU (*embarazable*); (2) según el aprendizaje con todas las variables vs. variables seleccionadas con FSS; (3)

según el uso o no de variables agregadas de los embriones; y, (4) según el umbral de decisión del clasificador probabilístico naive Bayes.

3.1. Predicción de embarazo

El primer conjunto de experimentos analiza los resultados de la metodología presentada para resolver el enfoque del ciclo de predicción del embarazo. En las Tablas 3.1 y 3.2 se muestran los resultados de los experimentos realizados.

3.1.1. Resultados sin variables agregadas

En la Tabla 3.1 se recogen los resultados de las pruebas utilizando únicamente las características de los ciclos recopiladas. Se han utilizado mediciones de recall y de precisión, que proporcionan información sobre la capacidad de predecir ejemplos positivos. Cabe destacar que los clasificadores NB-C que utilizan la selección de variables FSS (aquéllos que usan un subconjunto de variables consideradas relevantes y no redundantes) alcanzan valores de recall y de precisión mayores que los obtenidos por el clasificador NB-C aplicado al conjunto total de características recopiladas. De hecho, para el umbral de decisión 0.5, el recall de los clasificadores que realizan la técnica FSS indica que más del 40 % de los embarazos reales fueron identificados correctamente; y la precisión señala que aproximadamente el 60 % de los embarazos predichos son embarazos reales. En términos de la medida F, se observa que los modelos naive Bayes que se aprenden sobre el conjunto reducido de variables obtienen mejores resultados. En la Tabla A.1 se muestran las variables que forman el subconjunto reducido obtenido con el algoritmo de FSS.

3.1.2. Resultados con variables agregadas

Por otra parte, en la Tabla 3.2 se muestran los resultados de los experimentos utilizando las variables de ciclos junto a las variables agregadas construidas correspondientes a los embriones (Tabla 1.3). En comparación con los resultados de la Tabla 3.1, se observa que para los experimentos con el modelo naive Bayes sobre el conjunto completo de variables los resultados son ligeramente peores. Sin embargo, los resultados del clasificador NB-C aplicado al conjunto reducido que devuelve la técnica FSS son mejores en términos de la medida F. A medida que se rebaja el umbral, los resultados para el conjunto reducido son notablemente mejores que los obtenidos sin agregar las variables referidas a los embriones. Por lo tanto, se puede afirmar que las variables de los embriones, en formato agregado, aportan información relevante al desarrollo del ciclo y la capacidad predictiva de nuevos clasificadores. En la Tabla A.2 se muestran las variables que forman el subconjunto reducido obtenido con el algoritmo de FSS.

NB-C	Threshold	Recall	Precisión	F
NB	0.50	0.3639	0.3656	0.3648
	0.45	0.4131	0.4259	0.4194
	0.40	0.4824	0.4819	0.4821
	0.35	0.5264	0.5316	0.5290
	0.30	0.5831	0.5768	0.5799
	0.25	0.6195	0.6157	0.6176
FSS	0.50	0.4122	0.6034	0.4898
	0.45	0.5001	0.6118	0.5503
	0.40	0.6024	0.6197	0.6109
	0.35	0.6686	0.6270	0.6471
	0.30	0.6982	0.6317	0.6633
	0.25	0.7297	0.6423	0.6832

Tabla 3.1: Resultados de los clasificadores supervisados para la predicción de embarazo usando únicamente las variables de ciclos. Los datos de FSS representan los resultados del clasificador NB tras la selección de variables.

3.2. Predicción *embarazable*

El segundo conjunto de experimentos analiza los resultados de la metodología presentada para resolver el enfoque de la predicción de ciclos prometedores: la llamada predicción *embarazable*.

En este contexto de clasificación PU, se utilizan los datos no etiquetados y, por lo tanto, las métricas de evaluación estándar no son válidas. En las Tablas 3.3 y 3.4 se muestran los resultados de los experimentos realizados en términos del recall (proporción de instancias predichas como positivas en el subconjunto positivo) y F_{ps} [21]. Se han probado diferentes valores del parámetro p (la proporción desconocida real de ejemplos positivos en el subconjunto no etiquetado) para obtener buenos clasificadores. Para la validación se utiliza un esquema de 10-cv repetido 5 veces. En cuanto a la técnica FSS, se ha aplicado la adaptación al marco PU de la métrica CFS vista en la Sección 2.3.4 [23, 21].

3.2.1. Resultados sin variables agregadas

En la Tabla 3.3 se exponen los resultados de los experimentos utilizando únicamente las características de los ciclos recopiladas con un enfoque de clasificación PU. Se han utilizado mediciones de recall y la medida pseudo-F, que proporcionan información sobre la capacidad de predecir ejemplos positivos. Cabe destacar que los clasificadores que utilizan el conjunto reducido obtienen resultados similares a los obtenidos por el clasificador PNB aplicado al conjunto total de características recopiladas. Además, los valores del recall para el clasificador PNB indican que más del 70 % de los embarazos

NB-C	Threshold	Recall	Precisión	F
NB	0.50	0.3484	0.3915	0.3686
	0.45	0.3686	0.4036	0.3853
	0.40	0.3862	0.4192	0.4020
	0.35	0.3963	0.4342	0.4143
	0.30	0.4064	0.4487	0.4265
	0.25	0.4291	0.4655	0.4465
FSS	0.50	0.4836	0.5251	0.5035
	0.45	0.5478	0.5803	0.5636
	0.40	0.6082	0.6342	0.6209
	0.35	0.6560	0.6829	0.6692
	0.30	0.7177	0.7320	0.7248
	0.25	0.7618	0.7795	0.7705

Tabla 3.2: Resultados de los clasificadores supervisados para la predicción de embarazo usando las variables de ciclos y las agregadas de embriones. Los datos de FSS representan los resultados del clasificador NB tras la selección de variables

reales fueron identificados correctamente; aunque dado el valor de la métrica pseudo-F se puede intuir que la precisión del clasificador no será muy alta (el porcentaje de embarazos predichos que son reales no será muy elevado). Las variables que forman el subconjunto reducido de características obtenido por el algoritmo FSS se muestran en la Tabla A.3.

3.2.2. Resultados con variables agregadas

En la Tabla 3.4 se muestran los resultados de los experimentos utilizando las características de los ciclos recopiladas junto a las variables agregadas referidas a los embriones con un enfoque de clasificación PU. Se han utilizado mediciones de recall y la medida pseudo-F. En este caso, los clasificadores que utilizan el conjunto reducido obtienen resultados notablemente mejores que el clasificador PNB aplicado al conjunto total de características recopiladas junto a las agregadas. Tanto en términos del recall como de la métrica pseudo-F el clasificador construido con el conjunto reducido de variables obtiene mejores resultados. Por lo tanto, se espera que en términos de precisión también sea un buen clasificador capaz de predecir embarazos reales en un alto porcentaje. Por otra parte, los resultados en términos del algoritmo spy-EM son sensiblemente mejores que los resultados del algoritmo EM tradicional. Las variables que forman el subconjunto reducido de características obtenido por el algoritmo FSS se muestran en la Tabla A.4.

3.3. Conclusiones y trabajo futuro

En esta memoria se ha realizado un estudio del problema de las ART desde dos enfoques de modelado a partir de datos diferentes: uno de ellos mediante clasificación supervisada clásica y, un segundo, mediante clasificación positiva y no etiquetada, considerando también ejemplos de destino incierto para el aprendizaje del modelo. Para cada uno de estos enfoques se han desarrollado soluciones específicas que aprenden modelos de clasificación que aprovechan la mayor parte de los datos disponibles débilmente supervisados. De los resultados arrojados por las soluciones se pueden sacar varias conclusiones.

Los clasificadores aprendidos en el enfoque de predicción del embarazo de un ciclo han obtenido resultados notables según los datos de la Tabla 3.1. Los resultados con las variables de ciclo destacan que casi la mitad de los embarazos reales están correctamente identificados por los modelos aprendidos (recall ≈ 0.42) y, aproximadamente, el 60 % de los ciclos se predice correctamente como embarazo (precisión ≈ 0.6). Aunque son resultados para los que caben mejoras, son un buen punto de partida para afinar el proceso de aprendizaje de clasificadores. Asimismo, al incluir las características agregadas de los embriones los resultados muestran una ligera mejora en términos del recall y la métrica F, por lo que no se puede asegurar que los embriones no sean determinantes en el éxito de un ciclo de ART. La creación de variables agregadas con el conocimiento de los especialistas sería una línea interesante por la que seguir trabajando.

Por otra parte, en el enfoque alternativo de predicción *embarazable* se busca identificar ciclos prometedores, de forma independiente y previa al proceso de implantación. Como se explicó anteriormente, los casos fallidos no siempre pueden anotarse como ejemplos negativos según este punto de vista. La falta de ejemplos negativos en este planteamiento dificulta la comparación de los resultados con el enfoque de predicción del embarazo. Sin embargo, dado que la métrica pseudo-F es proporcional a la métrica F real, ésta se puede utilizar de forma fiable para la selección de modelos. Observando los resultados de las Tablas 3.3 y A.5, se puede concluir que los resultados de este enfoque en términos del recall, son significativamente mejores que los del enfoque de predicción del embarazo desde un punto de vista de clasificación supervisada. Incluso en el caso de $p = 0.75$, los clasificadores de la predicción *embarazable* muestran mayores valores de recall ($0.78 > 0.41$). Es decir, una gran proporción de los ciclos reales con pronósticos positivos se identifican realmente como positivos.

Otro de las perspectivas desde las que se comparan estos enfoques es desde el punto del aprendizaje con todas las variables o sólo aquellas seleccionadas con FSS. En las tablas de resultados se aprecia que en todos los casos el rendimiento de los modelos construidos con el conjunto reducido de variables son, como poco, similares a los obtenidos con el conjunto completo. Si bien en muchos casos se obtienen resultados mejores en términos de las

métricas calculadas. En este punto, con la ayuda de un especialista que seleccione buenas categorías para discretizar las variables antes de la construcción de los clasificadores se podrían conseguir resultados interesantes.

Por otra parte, en cuanto al uso de variables agregadas de los embriones, los resultados arrojan la idea de que puede ser beneficioso para el rendimiento de los clasificadores. Si bien no siempre se mejoran radicalmente los resultados, tampoco se empeoran en exceso. La creación de variables agregadas de información relevante con la ayuda de un experto podría ayudar a mejorar el rendimiento de nuevos clasificadores.

La solución propuesta evalúa la capacidad predictiva de la información recopilada por los médicos durante todo el procedimiento de ART. Con base en los resultados obtenidos, se puede afirmar que un sistema de recomendación para el problema del ART basado en los enfoques presentados proporcionaría información valiosa que podría implicar una mejora en la configuración de ciclos de buen pronóstico. Los clasificadores estudiados que predicen la viabilidad de un ciclo en el primer enfoque muestran un desempeño prometedor. Por otra parte, los resultados del enfoque de predicción *embarazable* también muestran resultados interesantes. En este sentido, sería interesante poder seguir aprendiendo modelos para estudiar su desempeño.

BNC	Threshold	Recall	F_{ps}
wPBC	0.50	0.5355	0.4897
	0.45	0.5316	0.4723
	0.40	0.4920	0.4485
	0.35	0.4831	0.4164
	0.30	0.4476	0.3612
	0.25	0.4191	0.3871
PNB	0.50	0.7030	0.5874
	0.45	0.7067	0.6168
	0.40	0.7185	0.6302
	0.35	0.7306	0.6737
	0.30	0.7459	0.7093
	0.25	0.7570	0.7028
EM	0.50	0.5209	0.3809
	0.45	0.5209	0.3809
	0.40	0.5172	0.3802
	0.35	0.5131	0.3764
	0.30	0.5109	0.3754
	0.25	0.5069	0.3735
spy-EM	0.50	0.5988	0.4795
	0.45	0.5788	0.4708
	0.40	0.5588	0.4608
	0.35	0.5451	0.4692
	0.30	0.5317	0.4579
	0.25	0.5181	0.4517
FSS	0.50	0.6830	0.5677
	0.45	0.6970	0.5618
	0.40	0.7083	0.5922
	0.35	0.7155	0.6374
	0.30	0.7381	0.6973
	0.25	0.7397	0.7081

Tabla 3.3: Tabla de resultados de los clasificadores PU únicamente para las variables de los ciclos. Los resultados mostrados para el clasificador PNB usan el parámetro $p = 0.55$ y los del algoritmo spy-EM se han construido con un nivel de ruido $l = 15\,\%$ y un porcentaje de espías $s = 10\,\%$.

BNC	Threshold	Recall	F_{ps}
wPBC	0.50	0.5312	0.4997
	0.45	0.5216	0.4826
	0.40	0.4786	0.4515
	0.35	0.4783	0.4181
	0.30	0.4665	0.3962
	0.25	0.4329	0.3847
PNB	0.50	0.5851	0.5757
	0.45	0.5888	0.5788
	0.40	0.5928	0.5827
	0.35	0.6007	0.5888
	0.30	0.6083	0.5950
	0.25	0.6083	0.5982
EM	0.50	0.5160	0.5884
	0.45	0.5160	0.5884
	0.40	0.5160	0.5884
	0.35	0.5160	0.5884
	0.30	0.5160	0.5884
	0.25	0.5160	0.5884
spy-EM	0.50	0.5835	0.6574
	0.45	0.5722	0.6583
	0.40	0.5701	0.6581
	0.35	0.5178	0.6580
	0.30	0.5669	0.6578
	0.25	0.5638	0.6577
FSS	0.50	0.6535	0.6374
	0.45	0.6572	0.6483
	0.40	0.6689	0.6501
	0.35	0.6938	0.6680
	0.30	0.7169	0.6719
	0.25	0.7358	0.6777

Tabla 3.4: Tabla de resultados de los clasificadores PU para las variables de los ciclos junto a las agregadas de los embriones. Los resultados mostrados para el clasificador PNB usan el parámetro $p = 0.55$ y los del algoritmo spy-EM se han construido con un nivel de ruido $l = 15\,\%$ y un porcentaje de espías $s = 10\,\%$.

Apéndice A

Tablas de resultados

Variable	Valores	Descripción
Tiempo Esteril	$(0, 3]$, $(3, 5]$, $(5, inf)$	Tiempo desde que se detectó la esterilidad
Emb. previos	Sí, No	¿Ha estado embarazada previamente?
Abo. previos	Sí, No	¿Ha sufrido algún aborto previamente?
Ciclos previos	0, 1, +2	Número de ciclos de ART a los que se ha sometido previamente
FSH	$(0, 5]$, $(5, 8]$, $(8, inf)$	Cantidad de hormona foliculoestimulante
lEnd	$(0, 9]$, $(9, 12]$, $(12, inf)$	Grosor endometrial
REM	$(0, 5]$, $(5, 15]$, $(15, inf)$	Recuento de espermatozoides móviles
TasaFertil	$(0, 0.5]$, $(0.5, 0.75]$, $(0.75, 1]$	Tasa de fertilidad (nEmbObten / MII)
nEmbTrans	0, 1, 2, 3	Número de embriones transferidos
transSelect	Sí, No	¿Se seleccionaron los embriones transferidos? (nEmbObten > nEmbTrans)

Tabla A.1: Tabla con el conjunto reducido de variables relevantes obtenido por la técnica FSS únicamente para las variables del ciclo en la predicción de embarazo.

Variable	Valores	Descripción
Tiempo Esteril	$(0, 3]$, $(3, 5]$, $(5, inf)$	Tiempo desde que se detectó la esterilidad
Emb. previos	Sí, No	¿Ha estado embarazada previamente?
Abo. previos	Sí, No	¿Ha sufrido algún aborto previamente?
Ciclos previos	0, 1, +2	Número de ciclos de ART a los que se ha sometido previamente
FSH	$(0, 5]$, $(5, 8]$, $(8, inf)$	Cantidad de hormona foliculoestimulante
TasaFertil	$(0, 0.5]$, $(0.5, 0.75]$, $(0.75, 1]$	Tasa de fertilidad (nEmbObten / MII)
nEmbTrans	0, 1, 2, 3	Número de embriones transferidos
numEmbCalA	0, 1, 2, 3+	Número de embriones de calidad A
numEmbZ1	0, 1, 2, 3+	Número de embriones de grado Z1
numEmb4cels	0, 1, 2, 3+	Número de embriones con 4 células
numEmbFragN1	0, 1, 2, 3+	Número de embriones de porcentaje de fragmentación $(0, 10]$

Tabla A.2: Tabla con el conjunto reducido de variables relevantes obtenido por la técnica FSS para las variables de ciclo junto a las agregadas de los embriones en la predicción de Embarazo.

Variable	Valores	Descripción
Tiempo Esteril	$(0, 3]$, $(3, 5]$, $(5, inf)$	Tiempo desde que se detectó la esterilidad
Emb. previos	Sí, No	¿Ha estado embarazada previamente?
Ciclos previos	0, 1, +2	Número de ciclos de ART a los que se ha sometido previamente
dEst	$(0, 9]$, $(9, 11]$, $(11, inf)$	Días de tratamiento de estimulación
FSH	$(0, 5]$, $(5, 8]$, $(8, inf)$	Cantidad de hormona foliculoestimulante
TasaFertil	$(0, 0.5]$, $(0.5, 0.75]$, $(0.75, 1]$	Tasa de fertilidad (nEmbObten / MII)
nEmbTrans	0, 1, 2, 3	Número de embriones transferidos

Tabla A.3: Tabla con el conjunto reducido de variables relevantes obtenido por la técnica FSS únicamente para las variables del ciclo en la predicción *embarazable*.

Variable	Valores	Descripción
Tiempo Esteril	$(0, 3]$, $(3, 5]$, $(5, inf)$	Tiempo desde que se detectó la esterilidad
Emb. previos	Sí, No	¿Ha estado embarazada previamente?
Ciclos previos	0, 1, +2	Número de ciclos de ART a los que se ha sometido previamente
FSH	$(0, 5]$, $(5, 8]$, $(8, inf)$	Cantidad de hormona foliculoestimulante
TasaFertil	$(0, 0.5]$, $(0.5, 0.75]$, $(0.75, 1]$	Tasa de fertilidad (nEmbObten / MII)
nEmbTrans	0, 1, 2, 3	Número de embriones transferidos
numEmbCalA	0, 1, 2, 3+	Número de embriones de calidad A
numEmbZ1	0, 1, 2, 3+	Número de embriones de grado Z1
numEmbFragN1	0, 1, 2, 3+	Número de embriones con porcentaje de fragmentación celular $(0, 10]$

Tabla A.4: Tabla con el conjunto reducido de variables relevantes obtenido por la técnica FSS para las variables del ciclo junto a las agregadas de embriones en la predicción *embarazable*.

PNB	Threshold	Recall	F_{ps}
$p = 0.33$	0.50	0.4224	0.4139
	0.45	0.4469	0.4214
	0.40	0.5288	0.4345
	0.35	0.5582	0.4997
	0.30	0.5815	0.5034
	0.25	0.6239	0.4934
$p = 0.45$	0.50	0.5918	0.4383
	0.45	0.6143	0.4413
	0.40	0.6371	0.4641
	0.35	0.6638	0.5089
	0.30	0.6792	0.5109
	0.25	0.7024	0.5259
$p = 0.55$	0.50	0.7030	0.5874
	0.45	0.7067	0.6168
	0.40	0.7185	0.6302
	0.35	0.7306	0.6737
	0.30	0.7459	0.7093
	0.25	0.7570	0.7028
$p = 0.65$	0.50	0.7614	0.6266
	0.45	0.7694	0.6346
	0.40	0.7773	0.6426
	0.35	0.7847	0.6506
	0.30	0.7963	0.6706
	0.25	0.8041	0.6998
$p = 0.75$	0.50	0.7824	0.6541
	0.45	0.7937	0.6622
	0.40	0.8089	0.6730
	0.35	0.8205	0.6827
	0.30	0.8281	0.6900
	0.25	0.8320	0.6963

Tabla A.5: Tabla de resultados de los clasificadores PNB según el valor del parámetro p únicamente para las variables de los ciclos.

spy-EM	Threshold	Recall	F_{ps}
$l = 5$	0.50	0.5749	0.4356
	0.45	0.5531	0.4375
	0.40	0.5293	0.4346
	0.35	0.5174	0.4352
	0.30	0.5040	0.4388
	0.25	0.4783	0.4367
$l = 10$	0.50	0.5462	0.4397
	0.45	0.5443	0.4410
	0.40	0.5304	0.4378
	0.35	0.5184	0.4372
	0.30	0.5045	0.4399
	0.25	0.4792	0.4386
$l = 15$	0.50	0.5988	0.4795
	0.45	0.5788	0.4708
	0.40	0.5588	0.4608
	0.35	0.5451	0.4692
	0.30	0.5317	0.4579
	0.25	0.5181	0.4517
$l = 20$	0.50	0.5954	0.4375
	0.45	0.5654	0.4391
	0.40	0.5337	0.4377
	0.35	0.5297	0.4345
	0.30	0.5261	0.4330
	0.25	0.5143	0.4359
$l = 25$	0.50	0.5485	0.4407
	0.45	0.5468	0.4392
	0.40	0.5368	0.4416
	0.35	0.5331	0.4407
	0.30	0.5212	0.4425
	0.25	0.5113	0.4423

Tabla A.6: Tabla de resultados de los clasificadores obtenidos con el algoritmo spy-EM para los diferentes niveles de ruido l únicamente para las variables de los ciclos. El porcentaje de espías usado es $s = 10\,\%$.

PNB	Threshold	Recall	F_{ps}
$p = 0.33$	0.50	0.4994	0.5012
	0.45	0.4994	0.5055
	0.40	0.4994	0.5099
	0.35	0.5031	0.5157
	0.30	0.5189	0.5272
	0.25	0.5298	0.5366
$p = 0.45$	0.5	0.5395	0.5155
	0.45	0.5551	0.5248
	0.40	0.5588	0.5295
	0.35	0.5588	0.5332
	0.30	0.5628	0.5383
	0.25	0.5628	0.5423
$p = 0.55$	0.50	0.5851	0.5757
	0.45	0.5888	0.5788
	0.40	0.5928	0.5827
	0.35	0.6007	0.5888
	0.30	0.6083	0.5950
	0.25	0.6083	0.5982
$p = 0.65$	0.50	0.6438	0.6159
	0.45	0.6438	0.6183
	0.40	0.6514	0.6233
	0.35	0.6514	0.6257
	0.30	0.6514	0.6274
	0.25	0.6514	0.6290
$p = 0.75$	0.50	0.6902	0.6751
	0.45	0.6939	0.6779
	0.40	0.6939	0.6791
	0.35	0.6976	0.6818
	0.30	0.6976	0.6832
	0.25	0.6976	0.6848

Tabla A.7: Tabla de resultados de los clasificadores PNB según el valor del parámetro p para las variables de los ciclos y las agregadas correspondientes a los embriones.

spy-EM	Threshold	Recall	F_{ps}
$l = 5$	0.50	0.5230	0.5865
	0.45	0.5170	0.5861
	0.40	0.5145	0.5845
	0.35	0.5136	0.5837
	0.30	0.5108	0.5825
	0.25	0.5101	0.5823
$l = 10$	0.50	0.5217	0.5857
	0.45	0.5188	0.5852
	0.40	0.5154	0.5845
	0.35	0.5137	0.5838
	0.30	0.5126	0.5837
	0.25	0.5121	0.5819
$l = 15$	0.50	0.5235	0.5874
	0.45	0.5222	0.5833
	0.40	0.5201	0.5819
	0.35	0.5178	0.5804
	0.30	0.5169	0.5787
	0.25	0.5138	0.5778
$l = 20$	0.50	0.5159	0.5873
	0.45	0.5149	0.5857
	0.40	0.5146	0.5844
	0.35	0.5137	0.5828
	0.30	0.5121	0.5812
	0.25	0.5105	0.5804
$l = 25$	0.50	0.5134	0.5825
	0.45	0.5132	0.5812
	0.40	0.5124	0.5764
	0.35	0.5113	0.5740
	0.30	0.5104	0.5732
	0.25	0.5067	0.5703

Tabla A.8: Tabla de resultados de los clasificadores obtenidos con el algoritmo spy-EM para los diferentes niveles de ruido l para las variables de los ciclos y las agregadas correspondientes a los embriones. El porcentaje de espías usado es $s = 10\,\%$.

Bibliografía

[1] S. Sunderam, D. M. Kissin, S. B. Crawford, S. G. Folger, D. J. Jamieson, L. Warner, and W. D. Barfield, "Assisted reproductive technology surveillance—united states, 2014," *MMWR Surveillance Summaries*, vol. 66, no. 6, p. 1, 2017.

[2] A. Villasante, L. Duque, and J. Garcia-Velasco, "Técnicas de reproducción asistida," vol. 3, pp. 199–204, 01 2005.

[3] J. Hernández González, "Contributions to learning bayesian network models from weakly supervised data: Application to assisted reproductive technologies and software defect classification," 2015.

[4] H. Achache and A. Revel, "Endometrial receptivity markers, the journey to successful embryo implantation," *Human reproduction update*, vol. 12, no. 6, pp. 731–746, 2006.

[5] T. Ebner, M. Moser, M. Sommergruber, and G. Tews, "Selection based on morphological assessment of oocytes and embryos at different stages of preimplantation development: a review," *Human reproduction update*, vol. 9, no. 3, pp. 251–262, 2003.

[6] L. Engmann, N. Maconochie, S. L. Tan, and J. Bekir, "Trends in the incidence of births and multiple births and the factors that determine the probability of multiple birth after ivf treatment," *Human Reproduction*, vol. 16, no. 12, pp. 2598–2605, 2001.

[7] C. M. Bishop, "Pattern recognition and machine learning, 2006," vol. 60, no. 1, pp. 78–78, 2012.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[9] S. A. Roberts, "Models for assisted conception data with embryo-specific covariates," *Statistics in medicine*, vol. 26, no. 1, pp. 156–170, 2007.

[10] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[11] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[12] C. Coughlan, W. Ledger, Q. Wang, F. Liu, A. Demirol, T. Gurgan, R. Cutting, K. Ong, H. Sallam, and T. Li, "Recurrent implantation failure: definition and management," *Reproductive biomedicine online*, vol. 28, no. 1, pp. 14–38, 2014.

[13] B. Calvo, I. Inza, P. Larrañaga, and J. A. Lozano, "Wrapper positive bayesian network classifiers," *Knowledge and information systems*, vol. 33, no. 3, pp. 631–654, 2012.

[14] B. Liu, W. S. Lee, P. S. Yu, and X. Li, "Partially supervised classification of text documents," in *ICML*, vol. 2, pp. 387–394, 2002.

[15] L. Scott, R. Alvero, M. Leondires, and B. Miller, "The morphology of human pronuclear embryos is positively related to blastocyst development and implantation," *Human reproduction*, vol. 15, no. 11, pp. 2394–2403, 2000.

[16] M. Ardoy, G. Calderón, G. Arroyo, J. Cuadros, M. Figueroa, R. Herrer, *et al.*, "Asebir criteria for the morphological evaluation of human oocytes, early embryos and blastocysts," *ASEBIR clinical embryology papers*, 2008.

[17] A. McCallum, K. Nigam, *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752, pp. 41–48, Citeseer, 1998.

[18] D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *European conference on machine learning*, pp. 4–15, Springer, 1998.

[19] Q. Yuan, G. Cong, and N. M. Thalmann, "Enhancing naive bayes with various smoothing methods for short text classification," in *Proceedings of the 21st International Conference on World Wide Web*, pp. 645–646, ACM, 2012.

[20] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[21] B. Calvo, "Positive unlabelled learning with applications in computational biology," *Department of Computer Science and Artificial Intelligence. San Sebastian, Spain: University of the Basque Country*, 2008.

[22] B. Efron, "Estimating the error rate of a prediction rule: improvement on cross-validation," *Journal of the American statistical association*, vol. 78, no. 382, pp. 316–331, 1983.

[23] M. A. Hall and L. A. Smith, "Feature subset selection: a correlation based filter approach," 1997.

[24] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[25] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, "Numerical recipes in c," *Cambridge University Press*, vol. 1, p. 3, 1988.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[27] F. Denis, A. Laurent, R. Gilleron, and M. Tommasi, "Text classification and co-training from positive and unlabeled examples," in *Proceedings of the ICML 2003 workshop: the continuum from labeled to unlabeled data*, pp. 80–87, 2003.