3 Aplicaciones del análisis de curvas ROC

3.1. El comienzo de las curvas ROC en Psicofísica

Los años 50, muy poco después de acabada la 2ª Guerra Mundial, es un periodo especialmente fértil en el intercambio de conocimiento entre diferentes disciplinas, con objeto de resolver problemas que no tenían solución desde un único campo. Muchos de estos problemas tenían que ver con la operación por las personas de sistemas complejos, algunos creados por el esfuerzo bélico, como fueron los radares, pero también otros muchos que exigían el estudio de las capacidades humanas mucho más allá de lo que se conocía en aquel momento.

Es en este periodo en el que se inician muchas áreas de conocimiento multidisciplinares, como por ejemplo la Ergonomía Cognitiva, o los Factores Humanos, como se denominó en los Estados Unidos. Es precisamente en los ambientes universitarios, científicos, pero también en los industriales (p.ej. Bell Labs) de aquel momento en los que se producen avances muy significativos que serán básicos para el desarrollo de lo que ahora es parte de nuestra vida cotidiana, como los sistemas informáticos.

Como producto de estos intercambios surge un conjunto de técnicas, conocidas como "análisis de curvas ROC" (acrónimo del inglés *Receiver*–o también *Relative - Operating Curve*), que recoge avances en tres áreas, la Teoría Estadística, la Psicofísica y la Teoría de Señales y los aplica a problemas de discriminación. Cada una de las áreas de conocimiento usará estas curvas de una manera u otra. El problema de la discriminación

es un clásico en la Psicología. De hecho, es uno de los problemas que abordó Fechner en sus "Elementos de Psicofísica", quizá el momento más claro de comienzo de la Psicología como disciplina científica. Ya desde un principio Fechner observa cómo uno de los principales problemas que se tienen que explicar es la enorme variabilidad en los umbrales –ya absolutos, ya relativos; su principal objeto de investigación.

Fechner propuso el método de las comparaciones binarias o de elección forzada para la búsqueda de un hipotético umbral psicofísico absoluto, o también relativo (la "diferencia apenas perceptible"). El principal problema de este constructo es su enorme variabilidad, debida a los muy diferentes procesos que separan la percepción propiamente dicha de la respuesta. "El principal problema es la gran variabilidad por diferencias individuales, tiempo, innumerables condiciones externas".

Bastantes años más adelante, la Teoría de Detección de Señales (TDS) se origina como una teoría matemática para explicar el proceso de detectar señales de radar (Peterson *et al*., 1954), pero se encontró que era muy útil para comprender el comportamiento de observadores humanos de señales visuales y auditivas simples (Tanner & Swets, 1954; Tanner, Swets, & Green, 1956).

El origen de las curvas ROC en la investigación psicofísica se puede situar al comienzo de la década de los 50, por Moncrieff Smith, Edna A. Wilson (Lincoln Laboratory del Massachusetts Institute of Technology), William A. Munson, John E. Karlin (Bell Telephone Laboratories), y Wilson P. Tanner y J. A. Swets (Universidad de Michigan). Se trata de un conjunto de estudios convergentes en los cuales se mostró que la predisposición de los observadores hacia una respuesta en particular (aparece el estímulo o no) es variable (hay observadores arriesgados o conservadores, o se puede conseguir que lo sean variando determinadas condiciones experimentales, por ejemplo, la matriz de pagos), y que al proceso de detección se añaden procesos de evaluación de probabilidades a priori y de costes o utilidades.

Por otro lado, en Teoría Estadística se desarrolla un procedimiento similar, la construcción de una prueba estadística. La distribución de ruido tiene su correlato en la distribución de la hipótesis nula, y la correspondiente al estímulo es la distribución de la hipótesis alternativa. De igual modo que la Teoría Psicofísica, la Estadística divide la

escala del eje x en dos regiones, mediante un criterio de decisión, o punto de corte (c) de tal modo que los valores muestrales de x menores que ese punto conllevan la aceptación de la hipótesis nula y los mayores conllevan la aceptación de la alternativa. El criterio o punto de corte se establece de tal modo que se minimiza el error de tipo I, esto es, el de aceptar la hipótesis alternativa cuando es la hipótesis nula la que es cierta.

El esquema completo de la decisión estadística, o la forma clásica del contraste estadístico de hipótesis, lo desarrollaron Jerry Neyman y Egon Sharpe Pearson en 1933. La forma común en estadística es fijar la probabilidad de error de tipo I de forma arbitraria (normalmente el famoso α=0.05), y elegir el criterio o punto de corte de tal modo que se minimice la probabilidad de un error de tipo II. Neyman y Pearson mostraron que la mejor prueba es la razón de verosimilitud.

Por otro lado tendremos la potencia del contraste. Normalmente se considera 1-β (1-prob de error II), y cuando se prueba la hipótesis nula frente a diferentes alternativas en vez de una sóla, la función de potencia tiene la misma forma que la función psicométrica ya definida por Fechner, y se puede demostrar que la regla de decisión es la misma que ya formuló Blackwell en su modelo.

Pues bien, la ROC, definida en Estadística, es la representación de 1 menos la función de potencia. La curva ROC es una forma de comparar dos características operativas, la típica del contraste estadístico, centrada en un enfoque conservador, fijando el error máximo permitido a la hora de rechazar la hipótesis nula, frente a una muy poco vista, como es la variación en la probabilidad de cometer un error tipo I fijando la probabilidad de cometer un error de tipo II. La curva ROC le da igual importancia a ambos tipos de error y muestra cómo varían conjuntamente según cambia el criterio para una diferencia dada entre las medias de las dos hipótesis.

Otro de los avances que contribuyeron a la propuesta de las curvas ROC fue el de la teoría de señales, y también una necesidad concreta que surge en aquellos momentos. Se trata de la decisión a partir de señales de radar, que mostró la importancia de disponer de criterios de decisión variables y de permitir diferentes reglas de decisión. Es en este contexto en el que surgen los conceptos ya clásicos de “falsas alarmas” (en estadística son los errores de tipo I) y “omisiones” (errores de tipo II). El propósito de variar las

reglas de decisión era tener en cuenta el contexto (situaciones defensivas u ofensivas) en las que variaban ampliamente los costes relativos y las ganancias (o pérdidas).

Su validez como teoría general aplicable a muchos campos de discriminación humana se apoyó en los descubrimientos empíricos en memoria de reconocimiento, y además, se planteó su aplicación a la capacidad discriminativa no ya sólo de los humanos, sino también de dispositivos o máquinas que ayudan o suplantan a los decisores humanos, como se ha mostrado en tareas de diagnóstico o en análisis de sistemas de recuperación de información (Swets, 1963, 1969, see Chapter 9). En todos estos estudios, el denominador común es un proceso de observación que proporciona grados variables de confianza sobre la ocurrencia de las alternativas que se han de discriminar y, segundo, un deseo de asignar esos grados varibles a una u otra alternativa de una manera razonable.

Como se recoge en cualquier manual de Psicología Básica, las curvas ROC se convierten en los años 50 y 60 en la piedra angular de la Psicofísica, y una simple búsqueda bibliográfica encontraría miles de publicaciones para casi cualquier dominio sensorial. Para hacerse una idea, en el prefacio a la reimpresión de su libro "*Signal Detection Theory and Psychophysics*", Green y Swets (original de 1966, reimpreso en 1988) presumen de recoger más de 1000 referencias en casi todas las dimensiones sensoriales. Una búsqueda bibliográfica en nuestros días producirá sin duda cantidades similares, si es que queda alguna dimensión sensorial para la que no se haya hecho el correspondiente estudio.

Pero nuestro interés viene por la aplicación de estas teorías y modelos a otro tipo de problemas: es en un conjunto de campos aplicados en los que el interés no es tanto en explicar los procesos subyacentes a las decisiones, sino de poder medir el nivel de agudeza discriminativa, la capacidad discriminativa o diagnóstica de un procedimiento o prueba, que será en muchos casos la base para una toma de decisiones (Swets & Pickett, 1982).

Una de las principales ventajas que aporta el modelo de la TDS, o de su pilar fundamental, las curvas ROC, es que proporciona un índice relativamente puro de

capacidad de discriminación, que es en gran medida independiente del criterio de decisión o de la tendencia hacia una elección u otra.

Por ejemplo, lo estricto que resulte el criterio usado en un sistema de diagnóstico por la imagen en Medicina clínica puede ser muy diferente en contextos de exploración, que en contextos de diagnóstico de confirmación. También en sistemas de predicción metereológica el criterio puede ser diferente entre diferentes regiones geográficas o de un usuario a otro.

En estos y otros contextos también se deben determinar las probabilidades de los diferentes resultados de la decisión, y en los beneficios y costes de estos resultados. Para estos contextos en los que las probabilidades, beneficios y costes son estables y se pueden estimar, el énfasis es más en el pago asociado con un punto particular de la ROC, esto es, con un criterio particular de decisión, más que en un índice para todos los puntos incluidos en la curva ROC.

Referencias esenciales de este largo camino son Green y Swets (1966), un libro esencial reimpreso por su gran demanda en 1988. Otras referencias importantes son D. McNicol (1972) y Egan (1975). En la Psicología son especialmente importantes las publicaciones de Swets en 1973 para *Science*, y el artículo de Swets en *Psychological Bulletin* en 1986. No entraremos, por tanto, en los temas de Psicofísica, que no por ser relativamente más antiguos no dejan de ser interesantes, pero que están claramente fuera del alcance de esta tesis doctoral.

Es curioso ver de una revisión de la literatura los caminos tan diferentes elegidos por Green y por Swets. El primero continuó sus trabajos más puramente psicofísicos (véase por ejemplo Green, 1993, Lee y Green, 1994, Gu y Green , 1994; quizá el más gracioso sea el de Forrest y Green, 1991, sobre la selección en moluscos...), sólo en una ocasión parece utilizar TDS para otros motivos fuera de Psicofísica -véase Fidell, Schultz y Green (1988) sobre tasas de prevalencia de las molestias inducidas por el ruido en poblaciones residenciales.

En nuestra búsqueda bibliográfica, los últimos artículos de Swets en Psicofísica son Swets y Birdsall (1978), en el que describe varios modelos de adaptación sensorial, o

atención progresivamente más focalizada, según TDS, y los de Getty, Swets, Swets y Green (1979) y Swets, Green, Getty y Swets (1978), que describen una serie de experimentos psicofísicos sobre la predicción de matrices de confusión obtenidas a partir de juicios de similaridad. Estos artículos, los últimos que firman Green y Swets conjuntamente, cuentan con otro Swets entre sus autores, Joel Swets, de quien nos preguntamos si es su hijo.

3.1 La dedicación de John A. Swets al análisis ROC para el diagnóstico

Parece ser 1978 el momento en el que J. A. Swets decide concentrarse casi exclusivamente al campo de aplicación de las técnicas TDS (especialmente el análisis de curvas ROC) al diagnóstico, sobre todo al diagnóstico por la imagen (sólo hay que pensar en el enorme desarrollo de técnicas como la radiografía, tomografía axial computerizada, etc. que exigen un entrenamiento específico de la persona elaborando el diagnóstico, protocolo que es sistemáticamente evaluado con técnicas basadas en curvas ROC).

Artículos pioneros en esta aplicación son los de Lusted (1971), McNeil y Adelstein (1976) y McNeil, Keeler y Adelstein (1975). Posteriormente Metz (1978) y Swets (1979), ambos en revistas médicas de este campo (*Seminars in Nuclear Medicine* e *Investigative Radiology*, respectivamente) se convierten también en referencias clásicas de esta primera etapa de aplicación del análisis ROC en los años 70. No debemos ignorar sin embargo el extrañísimo artículo de Swets para medir la eficacia de métodos de recuperación de información, ¡que fue publicado en 1969! No se vuelven a encontrar aplicaciones similares hasta mucho después, y aun todavía sigue siendo una metodología prácticamente desconocida (aunque veremos alguna que otra aplicación).

El papel de Swets en la evaluación de sistemas diagnósticos de casi cualquier tipo parece determinante. Y tampoco debemos olvidar el libro pionero de Swets y Picket (1982), así como el artículo previo en Science, en 1979, por Swets, Pickett, Whitehead, Getty, Schnur, Swets (¿su hijo otra vez?) y Freeman, y el de 1991 para *Medical Decision Making* (Swets *et al*. 1991), con otros 5 autores.

Siguiendo un poco en orden cronológico sobre las publicaciones del propio Swets, son reseñables las siguientes:

- En 1986, y para *Psychological Bulletin*, "*Indices of Discrimination or Diagnostic Accuracy: Their ROC and Implied Models*",
- en 1988, para *Science*, "*Measuring the Accuracy of Diagnostic Systems*",
- en 1992 para *American Psychologist*, "*The Science of Choosing the Right Decision Threshold in High-Stakes Diagnostics*",
- Humpreys y Swets (1991), en el que comparan dos estadísticos o dos procedimientos de evaluación: las correlaciones biseriales continuas y las ROCs para tres clases en la formación primaria de pilotos durante la Segunda Guerra Mundial. Las grandes muestras proporcionan una prueba severa de la aplicabilidad de las ROC a datos típicamente psicométricos. El modelo de la TDS ajustó los datos psicométricos bien. Ambos estadísticos están influidos por la restricción del rango pero la medida del área bajo la ROC (A_z) se desplazó menos rápidamente hacia su valor nulo que la correlación biserial (r_{bis}). La relación de r_{bis} a A_z es lineal según se modifica el rango de talento por cortes sucesivos en una distribución dada pero los dos estadísticos se separan cuando las distribuciones difieren: A_z es independiente de la forma de la distribución continua.
- Salter y Swets (1984) sobre las aplicaciones de la TDS a la metodología de investigación por encuestas.
- Getty, Swets, Pickett y Gonthier (1995), que parte de la pregunta ¿por qué los operadores humanos ignoran, o incluso desconectan, las señales de aviso? ¿Puede ser porque aparecen demasiado a menudo, y a los operadores les es demasiado difícil distinguir sobre la gravedad del aviso? Vemos este problema muy a menudo, por ejemplo, en la baja conciencia de riesgo y el desprecio a muchas señales de aviso en el caso de la conducción. Un aspecto esencial es que la probabilidad a priori de un evento como un accidente, en muchas situaciones cotidianas, es muy baja. Este interesante artículo estudia la relación entre el tiempo de respuesta de operadores humanos variando diferentes valores predictivos positivos y variando también los pagos por rendimiento del operador.

Y por último, curioso por lo estrambótico aunque reseñable,

- Swets y Bjork (1990), que revisan técnicas del ejército de EEUU para mejorar el rendimiento humano durante un estudio de 2 años. Se encontró poca evidencia de la efectividad de programas como gestión de estrés, o parapsicología. El resumen no explica el uso que se dio a las técnicas tipo "ROC".

Swets recibió el Premio de la APA a la distinguida contribución científica en 1990 (aparece recogido en *American Psychologist* en 1991) por sus contribuciones científicas, por sus trabajos en la TDS y su apliación a Psicofísica, además de haber demostrado la aplicabilidad de la percepción y cognición humanas a la evaluación de sistemas diagnósticos en una variedad de dominios.

Con cierto ánimo de escoger lo más representativo de la literatura sobre ROC, hemos llevado a cabo una revisión tan exhaustiva como ha sido posible de los artículos que utilizan o desarrollan la técnica ROC. Con este objeto hemos consultado la base de datos *PsycLit* de la APA, y los catálogos en línea disponibles en la Biblioteca de la UCM. Una búsqueda por “receiver operating”, sólo entre los años 2000 y 2004 produjo 205 entradas, incluyendo tesis doctorales, libros o capítulos de libros y artículos en revistas científicas. Una gran mayoría de estos artículos provienen de campos de la Psicología cercanos al campo médico o psiquiátrico, y especialmente destacan los estudios para comparar o proponer instrumentos de detección temprana. Este capítulo resume estas aplicaciones, sin entrar en detalles metodológicos, en cinco apartados, además del presente:

- A continuación lugar revisamos las aplicaciones al diagnóstico en Medicina y campos afines. Además de convertirse en un auténtico estándar de evaluación de indicadores clínicos, resulta muy interesante ver cómo se aplica en los ámbitos más avanzados de investigación bioquímica, y uno de ellos en particular se recoge en el apartado posterior dedicado a tecnologías de la información.
- En el tercer subapartado de este capítulo se reseñan artículos sobre la aplicación del análisis ROC a la detección temprana en Psicología Clínica, uno de los campos más productivos sobre este tema, quizá por su cercanía metodológica y temática con la Medicina.

- El cuarto apartado trata del análisis ROC en Psicología Social y Forense, entendiendo por ésta última las aplicaciones en ámbitos legales. Se tratan temas muy interesantes como la predicción de recaídas en maltratadores o en condenados en libertad condicional.
- El quinto apartado trata de la aplicación del análisis ROC a las tecnologías de la información.
- Por último un breve apartado recoge las escasísimas aplicaciones (menos una muy reciente) que hemos encontrado del análisis ROC en investigación de mercados propiamente dicha. Este hecho no nos desanima, sino que por el contrario nos motiva a proponer este conjunto de técnicas como muy interesantes y útiles para muchos de los propósitos de este área aplicada.

3.2 Aplicaciones del análisis ROC en Medicina

Sin ningún lugar a dudas, el campo científico en el que el uso del análisis de curvas ROC está más extendido es en la Medicina, en muchas de sus áreas de especialización. Unos protocolos diagnósticos precisos son parte crítica de cualquier sistema de salud. En Medicina, continuamente surgen nuevas pruebas diagnósticas o de detección temprana, y se deben evaluar empíricamente en función de su capacidad para discriminar los estados enfermos y no enfermo.

El área bajo la curva ROC ha adquirido rango de estándar como medida de precisión de una prueba diagnóstica, y se utiliza muy frecuentemente. Así por ejemplo, Rothman, Owens y Simel (2003) presentan un estudio para evaluar la eficacia diagnóstica de diferentes pruebas para la detección temprana de la otitis media aguda. Este artículo es representativo de muchos otros en los que resulta sorprendente ver la gran cantidad de información que se ofrece en el mismo resumen, con las áreas bajo la curva ROC y su intervalo de confianza al 95%, que permite al lector hacer una comparación muy precisa de la capacidad diagnóstica de los indicadores bajo prueba desde el principio de la lectura.

Hay una corriente en la Medicina denomina "*Evidence-based medicine*" (Jenicek, 2003), que insiste en el aspecto metodológico de las pruebas diagnósticas. Jenicek

dedica en su libro apartados a medidas de las pruebas: sensibilidad, especificidad, valores predictivos, razones de verosimilitud, ROC, ratios de riesgo, eficacia, efectividad y eficiencia. Algunos otros conceptos relacionados en Medicina son los de diagnóstico diferencial y meta-análisis, y se complementan con aplicaciones de teoría del caso y lógica difusa, árboles de decisión y requisitos estructurales.

Numerosos artículos han divulgado las técnicas de análisis ROC en campos normalmente relacionados con diagnóstico:

- Hsiao, Bartko y Potter (1989) proporcionan una breve recopilación no matemática de los principios subyacentes a ROC.
- Fombonne (1991) introduce los principios básicos del análisis ROC, y se centra en cómo las curvas ROC pueden ayudar para la selección de un punto de corte óptimo, que se ha mostrado que depende de la tasa de prevalencia, de las consecuencias de las clasificaciones correctas e incorrectas, y de las distribuciones de las escalas. Se utiliza el *Child Behavior Checklist*, y se discute sobre la aplicación de diferentes puntos de corte.
- Fombonne y Fuhrer (1992) publican otro artículo describiendo las bondades de ROC, especialmente frente a las típicas pruebas t de Student.
- Somoza y Mossman (1991a) describen el conjunto de supuestos matemáticos que se pueden utilizar para enlazar las curvas ROC con las distribuciones subyacentes de valores de la variable diagnóstica bajo medición. Estos supuestos se ilustran con una prueba diagnóstica que distingue abusadores de alcohol de los que no lo son.
- Mossman y Somoza (1991) en este otro artículo muy relacionado con el anterior, proponen que los clínicos pueden ganar una mejor compensión del papel de las pruebas diagnósticas después de conocer bien las descripción de eficacia proporcionadas por ROC.
- Somoza y Mossman (1991b), en el tercer artículo relacionado en este mismo año, describen una técnica matemática para operacionalizar y evaluar las pruebas diagnósticas que combina la TDS y la teoría de la decisión basada en la utilidad. Esta técnica se utiliza para datos de 4 estudios en los cuales la arquitectura del sueño se utiliza como marcador biológico para la depresión. Los resultados muestran cómo las utilidades de los resultados influyen en el punto de corte óptimo para la latencia

REM, y cómo esta relación se ve influida por la prevalencia de la depresión en la población bajo prueba. Se realizan cálculos de los límites prácticos que deben imponerse en las incertidumbres sobre utilidades para operacionalizar una prueba diagnóstica para una situación clínica específica.

- Mossman y Somoza (1989) publican en este estudio un análisis ROC reevaluando datos de 7 estudios anteriores.
- Harber (1981) revisa las medidas de toma de decisiones aplicadas en educación especial, y muestra rationale y procedimiento para ROC en este campo. Habla de precisión diagnóstica, sensibilidad y especificidad.

Dos autores destacan entre todos los que aparecen en la literatura de análisis ROC en Medicina: Metz y Hanley. Recordemos que Metz es pionero en la aplicación del análisis ROC al diagnóstico por la imagen con su trabajo de 1975 (Metz, Starr, Lusted y Rossman, 1975) y el artículo de 1978 en *Seminars in Nuclear Medicine*, y continúa su trabajo tanto de desarrollo metodológico -que aparecerá reflejado adecuadamente en los capítulos 4 y 5- como de divulgación (Metz, 1986).

Los trabajos de Metz también aparecen en el campo psicológico, pero especialmente en sus aspectos más metodológicos: Metz y Kronman (1980) proponen pruebas de significación estadística para comprobar diferencias entre ROCs empírica y esperada binormal, entre 2 ROC binormales independientes, y entre grupos de binormales independientes.

Por otro lado, Hanley es también autor de un conjunto importante de artículos tanto de divulgación en varias áreas médicas, como del desarrollo metodológico del análisis ROC. Sobre el primer aspecto destacar Hanley y McNeil (1982), Hanley (1989), y un artículo básico para entender el avance metodológico realizado en el campo del diagnóstico médico, el artículo clásico de Hanley y McNeil (1983) en el que proponen un método para comparar las áreas bajo diferentes curvas ROC derivadas de los mismos casos. Otro artículo clave en este momento es el de McNeil y Hanley (1984), en el que explicitan en su título la consideración de las curvas ROC como una herramienta estadística.

El uso del análisis ROC está muy extendido en el campo de la Medicina, incluyendo aspectos de gestión de salud pública y centros de asistencia primaria: Braga y Oliveira (2003) recomiendan el análisis de curvas ROC, y en particular el área bajo la curva, como una herramienta muy poderosa para medir y especificar los problemas de rendimiento diagnóstico en Medicina. En su ejemplo muestran la aplicación de estas técnicas para definir cuál de entre cinco índices de severidad clínica se puede considerar el mejor para evaluar el riesgo de muerte para niños recién nacidos con muy bajo peso, todos de unidades de cuidados intensivos de hospitales portugueses. En este primer estudio se utilizan muestras correlacionadas, y en un segundo estudio utilizan muestras independientes de cuatro hospitales portugueses, con el objeto de identificar la unidad de cuidados intensivos con mejor rendimiento, a partir de los índices de severidad clínica.

Zygowicz y Saunders (2003) desarrollan una medida para la detección temprana de problemas conductuales para su uso con adultos jóvenes en servicios de salud primaria. Sostienen que estos cuestionarios de salud mental se debieran incorporar por rutina en los cuidados médicos para ayudar en la identificación de la enfermedad mental en jóvenes adultos. Se utilizaron curvas ROC para calcular el punto de corte que permita distinguir mejor entre dos muestras, una de 134 pacientes y otra de 233 de grupo control.

La aplicación de estas técnicas de evaluación de sistemas diagnóstico permite la comparación de enfoques o protocolos médicos, no sólo para diferentes enfermedades sino también entre diferentes profesionales, países, etc. Por ejemplo, Esserman *et al*. (2002) compara enfoques centralizado de mamografía (p.ej., Reino Unido y Suecia) con énfasis en alta especificidad y alta sensibilidad. Por contraste, EEUU no tiene este enfoque centralizado, y enfatiza alta sensbilidad, pero tiene en promedio más baja especificidad.

El área de diagnóstico por la imagen, quizá por la influencia temprana de Swets en este campo, ha sido el que más ha utilizado las técnicas de análisis de curvas ROC. En Radiología se utilizan las curvas ROC para modelizar y evaluar el rendimiento de radiólogos. Estos datos se pueden utilizar para optimizar la detección temprana de cáncer mediante mamografía.

Otra de las aplicaciones en Medicina es la mejora de los indicadores, del conjunto del sistema de predicción. Por ejemplo, Schaffer *et al*. (2001) evaluá la eficacia de programa para buscar en una base de datos de proteínas mediante análisis de curvas ROC.

Otras aplicaciones que se pueden mencionar son:

- evaluar indicadores directos versus indirectos de calidad de hospital para niños prematuros (Rogowski *et al*. 2004),
- el valor de prognosis del factor de crecimiento de la placenta en pacientes con dolor agudo de pecho (Heeschen *et al*. 2004), y
- la predicción de mortalidad entre pacientes hospitalizados por fallo coronario (Lee *et al*. 2003). En este último estudio se señala la *Odds-ratio* (OR) como indicador básico de relación diagnóstica, y el análisis de curvas ROC se realizó mediante STATA v. 7 y la optimización de las mismas se realizó utilizando S-plus.
- desarrollo de instrumentos de detección rápida: para la detección de la epilepsia (Barr *et al*. 2004), la estancia hospitalaria (Holland *et al*., 2003), la detección del síndrome de ansiedad generalizada (Fresco *et al*., 2003), el diagnóstico temprano de la demencia (Jager *et al*, 2003), la fobia social (Newman *et al*., 2003)
- Hay *et al*. (2004) validan dos escalas para el diagnóstico del desorden de coordinación durante el desarrollo (DCD). Se utilizaron las curvas ROC y el estadístico kappa para la evaluación del rendimiento y la estimación de sensibilidad y especificidad de la prueba.

Un estudio prototipo es el de Clarket, McKenzie y Dean (1994). Estos autores probaron la hipótesis de que, habiendo determinado un punto de corte óptimo y para un instrumento de *screening*, o para el uso en la práctica diagnóstica en psiquiatría, no todas las combinaciones de los items rinden igualmente bien. Se obtuvieron datos de 154 pacientes no psiquiátricos que habían compeltado el *General Health Questionnaire* y el *Structured Clinical Interview for Diagnostic*. El programa CUTOFF, empleando la metodología *Quality ROC Curve*, fue usado para encontrar el punto óptimo de corte de 7/8 para detectar la depresión. Consistente con las hipótesis, los resultados demuestran que para un punto de corte de y items en el instrumento de detección, hay gran

variabilidad en la identificación de las diferentes combinaciones de y items. El uso de puntos de corte no es quizá tan fiable como se piense algunas veces.

Un estudio muy interesante es el de Goodman, Slap y Huang (2003), en el que evalúan indicadores socioeconómicos como predictores de depresión y obesidad en la adolescencia. Este estudio, por otro lado típico de la epidemiología, trata de las tasas de prevalencia de obesidad y de los predictores socio-económicos (*SES o Socio-Economic Status*) para predecir el parámetro PAR (*Population Attributable Risk*).

La población geriátrica es también un objetivo muy frecuente de estas pruebas de detección rápida. Leahy *et al*. (2003) utilizan una herramienta como para detectar incapacidades funcionales en población geriátrica, para lo que utilizan varios puntos de corte que proporcionan sensibilidades "entre 60 y 100%". Ruckdeschel *et al*. (2004) tienen como objetivo la detección de la depresión en personas en residencias geriátricas. Los factores predictivos de la demencia son el objetivo del trabajo de Van Hout *et al*. (2002), y específicamente, para el desarrollo de recomendaciones para el trabajo diagnóstico en médicos de familia. Los resultados muestran que las actividades de la vida diaria, el número de años desde que los síntomas comenzaron, y la presencia de otros problemas somáticos contribuyen a la predicción de la presencia o ausencia de demencia. Se calculó el área bajo la curva ROC con estas tres variables y se obtuvo un área de 0.79. En el caso de la diagnosis de los médicos de familia era de 0.74, con lo que concluyen que la capacidad predictiva de estos últimos es razonable.

Hay varios estudios dedicados a evaluar la capacidad diagnóstica de una prueba conocida como "del reloj" para la demencia senil. Storey *et al*. (2002) evalúan la capacidad diagnóstica en una población multicultural, de habla no inglesa, en particular con 93 pacientes de 78 años de media. Cada paciente era evaluado por un geriatra que recogía datos demográficos, administraba otras medidas, y categorizaba cada paciente como normal o demente. Cada dibujo de reloj se puntuaba de acuerdo con una escala de 4 puntos denominada CERAD. Se encuentran sensibilidades del 16% pero a costa de especificidades del 16%. Sólo uno de los métodos consiguió una especificidad por encima del 50%, pero con una sensibilidad intermedia del 78%. Estos autores concluyen que no hubo diferencias significativas en los métodos de puntuación para detectar la demencia, y que la prueba del dibujo del reloj fue en el mejor de los casos, de modesta

capacidad predictiva, con baja especificidad en todos los métodos. Esta prueba puede no ser un buen disciminante para toda la población. Sin embargo, un poco antes, para esta misma prueba y al menos el primer autor, Storey *et al*. (2001) obtuvieron que la precisión diagnóstica en pacientes hablantes de inglés, a partir de las áreas bajo la curva ROC fue de 0.79. , con ligeras variaciones entre los 5 métodos diferentes de puntuación que había disponibles.

El campo médico es también muy proclive a probar los últimos modelos de minería de datos, como las redes neuronales. Véase Das *et al*. (2003) como un ejemplo de su aplicación para la prognosis de la hemorragia gastrointestinal, siempre comprobando la validez interna y externa de su modelo predictivo. En particular compara métodos de redes neuronales y regresión logística para la predicción de la hemorragia gastrointestinal. Las curvas ROC se utilizan para comparar la eficacia de los modelos, así como sus índices de precisión total, sensibilidad, especificidad, valor predictivo positivo (PPV), valor predictivo negativo (NPV).

Las aplicaciones de curvas ROC trascienden las áreas de mayor relevancia clínica para adentrarse en técnicas de investigación bioquímica, como evaluar indicadores genéticos como predictores (Parodi *et al*. 2003; Pepe, Longton, Anderson y Schummer, 2003).

Este último estudio es especialmente interesante y está encabezado por una de las autoras más prolíficas en este campo, Margaret S. Pepe. En él se proponen medidas basadas en curvas ROC para la distinción entre tejidos cancerosos o normales. También proponen que la variabilidad muestral en las ordenaciones de genes se puede cuantificar, y sugieren utilizar la "función de probabilidad de selección", la distribución de probabilidad de ordenaciones de cada gen, estimada por medio de *bootstrap*. Se utilizan estudios de simulación para evaluar el rendimiento relativo de las diferentes medidas de ordenación de genes y la cuantificación de variabilidad muestral.

En el artículo señalan que para el propósito de la discriminación entre estos genes, tradicionalmente se han utilizado técnicas de regresión para la identificación de combinaciones de genes que proporcione discriminación entre los diferentes tipos, mientras que para desarrollar nuevas clasificaciones de tipos se han utilizado técnicas de *cluster*.

Un problema de la Medicina, al igual que la Psicología y otras ciencias de la salud o de la conducta, es la gran cantidad de variables que se manejan en los modelos. Qu, Adam, *et al*. (2003) proponen un método de reducción de datos utilizando la "*Discrete Wavelet Transform*" para el análisis discriminante de datos con muy alta dimensionalidad.

El campo médico ha sido en el que más se ha avanzado no sólo en la aplicación práctica de curvas ROC a problemas específicos, sino también en el desarrollo metodológico, como muestran diferentes artículos que luego recogeremos más en detalle (capítulo 4). Así, Eguchi y Copas (2002) presentan para su estudio varios ejemplos incluyendo un estudio de diagnosis médica en citología de pecho, Thompson (2003) y Dodd y Pepe (2003a) presentan ejemplos de aplicación de detección temprana del cáncer de próstata.

Otra de las aplicaciones que se ha producido en Medicina es el meta-análisis. Dukic y Gatsonis (2003) proponen un nuevo método meta-analítico para evaluar la precisión de una prueba y llegar a una curva ROC resumen para una colección de estudios que evalúen pruebas diagnósticas. Se discuten formulaciones bayesianas y no bayesianas del enfoque. En el bayesiano se proponen varias formas de construir curvas ROC resumen y sus bandas creíbles. Se ilustra el enfoque con datos de un meta-análisis de progesterona para diagnóstico de embarazo.

Shang, Lin y Goetz (2000) comparan la capacidad predictiva de modelos de diagnosis de MRSA (*Methicilin-Resistant Staphylococcus Aureus*), mediante regresión logística y red neuronal, mediante curvas ROC y método de validación cruzada. Realizan contraste estadístico de las áreas bajo la curva según se muestra en la figura inferior. Finalmente se encuentran redes neuronales mejores, más robustas y potencia discriminatoria.

3.3 El análisis ROC y la detección temprana en Psicología Clínica

Lógicamente, el campo de mayor aplicación del análisis de curvas ROC en Psicología es el más próximo a la Medicina, por supuesto si excluimos todas las aplicaciones todavía vigentes de las curvas ROC en Psicofísica. Resulta paradójico este efecto de

"ida y vuelta" entre dos disciplinas haya generado una riqueza enorme en las posibilidades de un conjunto de métodos, por lo demás, bastante sencillos.

La detección temprana (quizá sea preferible su acepción inglesa, tan económica, pero tan significativa, "*screening*") para multitud de problemas es un objetivo de cualquier sistema diagnóstico, y es un tema clásico en la literatura del análisis de curvas ROC. Desde hace mucho tiempo se ha encontrado en el análisis ROC la herramienta básica para compara la eficacia de distintos procedimientos y para mejorar su aplicación. Grossberg y Grant (1978) presentan en un artículo pionero sobre cómo la TDS y los procedimientos de estimación de magnitudes se están aplicando a una variedad creciente de problemas antes recalcitrantes en Psicología clínica y en toma de decisiones médicas. Uno de los ejemplos que presentan es la evaluación del dolor, en la que un número de procedimientos o manipulaciones puede afectar la intención de los sujetos para informar del dolor, pero dejan inafectadas la detectabilidad de los estímulos que producen dolor.

Grossberg y Grant señalan que estos métodos muestran también superior acuerdo con medidas psicofisiológicas de esos factores y sugieren que ambas técnicas representan un avance significativo sobre prácticas actuales en términos de su mayor objetividad y precisión, uso parsimonioso de un lenguaje único para laboratorio y clínica, y potencial para la cuantificación de conductas sutiles o encubiertas. Entre los ejemplos destacan aplicaciones de ambas técnicas para evaluación y comprensión de dolor, ansiedad, drogas psicoactivas y toma de decisiones médicas.

La importancia de estas técnicas de detección temprana y la evaluación de su capacidad predictiva motivan extensas recopilaciones como la de Mc Fall y Treat (1999) para el *Annual Review of Psychology*, en la que, no obstante, recogen muy pocas de las publicaciones más metodológicas que se recogen en esta tesis.

McFall y Treat hablan de la necesidad de "medir las medidas", y proponen para ello los esquemas bayesiano y de la TDS. Aunque su revisión resulta en algunos puntos decepcionantemente superficial, intenta ser didáctico y para ello propone un ejemplo de Somoza *et al*. (1994) sobre pruebas psicológicas de depresión. En un punto, estos autores resaltan que "una búsqueda sistemática de la literatura empírica en la Psicología

Clínica mostró sorprendentemente pocos ejemplos publicados de aplicaciones a problemas clínicos reales". Entre ellos señalan:

- El de Camasso y Jagannathan (1995) para detectar la presencia de maltrato infantil,
- Dalgleish (1988) sobre toma de decisiones sobre acogimiento infantil,
- Erdman *et al*. (1987) sobre la predicción de intentos de suicidio,

y algunos otros que reseñaremos posteriormente como parte de nuestra revisión.

McFall y Treat señalan algunos factores que han dificultado la difusión de estas técnicas (tanto las bayesianas como las de análisis ROC) en la Psicología. En primer lugar, las demandas cuantitativas de las técnicas de análisis ROC. En segundo, las demandas conceptuales que "pueden ser intimidantes para psicólogos clínicos cuya formación cuantitativa ha estado limitada a los cursos tradicionales de estadística durante los estudios de Psicología". Por algún motivo, señalan, "este modo de pensar no parece establecerse". Habrá que preguntarse por qué en Medicina sí y por qué en Psicología no. Puede ser revelador un artículo de Meehl al respecto (Meehl, 1978).

Otra crítica que recogen en su revisión es la de los supuestos de la TDS, que pueden no ser apropiados para el campo de aplicación. En su forma original, por ejemplo, la TDS asumía que la variable utilizada para discriminar entre los dos estados a detectar fuera normal en cada estado. Como ya sabemos en muchos campos, este supuesto no es válido. ¿Pero es que tampoco es válido en la Medicina?

Otra crítica recogida: su limitación a problemas de diagnóstico con sólo decisiones dicotómicas. Aunque señalan que éste es un problema menor: cualquier clasificación o decisión puede ser dicotomizada (y si se permite el comentario, el caso más frecuente es que efectivamente sea dicotómica).

Se hacen eco de una limitación expuesta por Swets (1988) en relación con la fiabilidad y validez del análisis ROC como método de cuantificar el valor de la información y la precisión de las pruebas diagnósticas. Una limitación es el llamado problema del "estándar dorado" (*golden standard*). "Si no podemos determinar con certeza para cada caso en nuestra muestra el estado verdadero, esto es, si cada caso es positivo o negativo,

entonces no podemos posiblemente esperar que la TDS nos proporcione una evaluación válida de la precisión de la prueba". Y pone el siguiente ejemplo: "¿Cómo se pueden determinar la potencia discriminatoria de las pruebas de polígrafo en casos reales de crímenes si la verdadera culpabilidad o inocencia no se puede establecer con certeza?" Otro problema ocurre cuando el sistema de evaluación y la determinación de la verdad no son independientes. Por ejemplo, "si el *gold standard* se establece en los casos criminales por las confesiones de los criminales, y la prueba del polígrafo se usa para predecir la culpabilidad o la inocencia, entonces el propio sistema predictivo puede contaminar la verdad porque las confesiones pueden ser más probables después de que el polígrafo ha encontrado la culpabilidad". Este es un problema también cuando los procedimientos para determinar el "*gold standard*" influyen en la selección de casos para la muestra de prueba. En cualquier caso, Swets señala que todos estos problemas surgen de las debilidades o fallos en el diseño de nuestras pruebas o en nuestra incapacidad para determinar la verdad.

Con cierto entusiasmo, estos autores señalan que no existía excusa en ese momento (1999) para no utilizar las técnicas de la TDS, puesto que estaban totalmente accesibles (cita la recopilación de obras de Meehl, 1973, y de Swets, 1996) pero por ejemplo no cita ninguno de la gran cantidad de excelentes introducciones o artículos didácticos sobre el tema en Medicina.

Algún efecto ha tenido que tener esta revisión y estas recomendaciones, porque en nuestra revisión hemos encontrado bastantes aplicaciones en Psicología Clínica, que revisaremos a continuación.

A modo de conclusión parcial veremos cómo una mayoría de estudios se basan en la aplicación de técnicas de regresión logística y la aplicación de curvas ROC tanto para la comparación de modelos o instrumentos entre sí, como con respecto a un estándar o capacidad pre-establecida, y algunos utilizan también las posibilidades de esta metodología para mejorar u optimizar la capacidad predictiva variando los umbrales de decisión. Todos estos aspectos, desde el punto de vista de la metodología, se recogen con detalle en los capítulos 4 y 5 de este trabajo.

3.3.1 La detección temprana de trastornos psicológicos

Por ejemplo, Edens, Buffington, Tomicic y Riley (2001) utilizan PPI (*Psychopathic Personality Inventory*) como detección temprana de psicopatía. Se utilizó un diseño análogo de medidas repetidas para que 186 respondientes completaran el PPI bajo dos condiciones y con instrucciones específicas para crear una impresión favorable de ellos mismos. En la condición "bueno falso", los participantes pudieron aparecer significativamente menos psicopáticos, en compración con los que tuvieron mayores puntuaciones en las condiciones estándar. Los análisis de curvas ROC indicaron que aunque una de las escalas diferenciaba significativamente entre los buenos falsos y los honestos (área bajo la curva de 0.73), se producía un número considerable de malas clasificaciones.

La detección de la depresión, aislada o en conjunto con problemas de abuso de drogas o alcohol, que veremos más adelante, es quizá uno de los campos de más frecuente aplicación de las técnicas de análisis ROC en Psicología Clínica:

- Timbremont, Braet y Dreessen (2004) examinaron la utilidad del *Children's Depression Inventory* (CDI) para predecir un diagnóstico de depresión según la DSM-IV. Se utilizó un enfoque categórico mediante análisis de curvas ROC para evaluar la adecuación de puntuaciones de corte con propósitos de detección temprana. Los resultados indican un punto de corte sugerido entre 13 y 19, y un punto de corte de 16 mostró una relación óptima entre sensibilidad y especificidad.
- Viinamaki *et al*. (2004) utilizan el *Beck Depression Inventory* para detectar la depresión en diferentes fases.
- Muller *et al*. (2003) utilizan la *Hamilton Depression Rating Scale* para detectar la depresión, y realizan curvas ROC, obteniendo un punto de corte de 31, para una sensibilidad del 93.5% y una especificidad del 83.3%.
- Zimmerman *et al*. (2004) utilizan un cuestionario breve, con buenas propiedades psicométricas para detectar los desórdenes más comunes en el eje I de la DSM-IV, que se pueden encontrar en contextos de salud mental. Se examinó el rendimiento de este cuestionario en pacientes ambulatorios con dependencia o abuso de drogas y alcohol, y se determinó si su rendimiento en pacientes con este perfil es tan bueno como en un perfil sin abuso de sustancias. Se usaron las curvas ROC para cada

subescala y se muestra que todas las áreas bajo la curvas son significativas (diferentes de 0.5 o predicción por azar) y similares en los dos grupos.

- La validez comparativa de diferentes cuestionarios de detección temprana para desórdenes depresivos según la DSM-IV y los diagnósticos de los médicos es el objetivo de estudio de Lowe *et al.*, (2004). Destacan que uno de ellos (PHQ) es significativamente superior a los otros, aunque "para cualquier síndorme depresivo" "la diferencia total no alcanzó significación estadística al 5%". Este artículo recomienda puntos de corte para obtener sensibilidades entre el 98% (PHQ) o inferiores, para los otros instrumentos de detección.
- Dendukuri, McCusker y Belzile (2004) evalúan la validez de una herramienta de detección rápida de problemas funcionales y depresión, y para predecir un incremento en el uso de servicios de salud en paciente de más de 65 años, que estaban a punto de ser dados de alta en un servicio de emergencia. La validación de criterios de línea base incluyó el estado funcional premórbido en ambos estudios y de la depresión sólo en el segundo. Se estimaron las áreas bajo la ROC para la validez concurrente de la escala, y se obtuvo entre 0.65 y 0.86.

Una búsqueda bibliográfica para periodos anteriores a 1998 muestra una menor frecuencia de aplicaciones del análisis de curvas ROC en Psicología Clínica, pero todavía sigue habiendo un número importante. Algunas de las más significativas se resumen en la tabla 3.1.

3.3.2 Predicción del uso de los servicios de salud

Las aplicaciones se extienden a problemas complejos de administración de todo el sistema sanitario: Rakovski *et al*. (2002) intentan determinar qué información, tomada de los datos administrativos, permiten predecir, o identificar las personas que más probablemente serán usuarios del sistema de salud en el año siguiente. La variable criterio era la utilización, en días (1-365), del sistema sanitario, y se dicotomizó utilizando 92 días como umbral (esto es, estar entre el 2% superior o no). Utilizan tres tipos de modelos de regresión logística: (1) modelos de uso previo (uso en año anterior y edad y sexo), (2) modelos diagnósticos propios de la Administración Americana (HCC y ADG), además de edad y sexo, y (3) modelos combinados, esto es, el uso

previo además de los modelos diagnósticos. A través del análisis de curvas ROC y los resultados de clasificación de los 3 modelos, Rakovski *et al*. encuentran que el mejor modelo predictivo es el mixto.

Tabla 3.1. Resumen de aplicaciones del análisis de curvas ROC en los años 90 o antes

Camasso y Jagannathan (1995)	Comparan la capacidad predictiva de las pruebas psicológicas *Illinois* CANTS 17B y WSRM (*Washington State Risk Matrix*)
Olin, John y Mednick (1995)	Utilizan cuestionario de 25 elementos completado por profesores para la detección de esquizofrenia en Copenhague.
Vida, Des Rosiers, Carrier, Gauthier (1994)	Utilizaron el análisis de curvas ROC para comparar la eficacia de la *Escala Cornell para Depresión en Demencia* (CSDD) y la *Escala Hamilton para Depresión* (HRSD) para detectar criterios diagnósticos de depresión aguda en sujetos con Alzheimer.
Chen, Faraone, Biederman y Tsuang (1994)	Se examinaron las capacidades (precisión) diagnósticas de dos instrumentos para el desorden de déficit de atención por hiperactividad, mediante el *Checklist* de conducta de niños (*Child Behavior Checklist* - CBCL).
Somoza, Steer, Beck y Clark (1994)	Evaluaron 3 instrumentos: *Beck Depression Inventory Revised*; la subescala de depresión de un *checklist* cognitivo desarrollado por A. T. Beck (1987); y la *Hamilton Rating Scale for Depression*. La *Hamilton Rating Scale for Depression* fue superior a las otras escalas de depresión.
Draijer y Boon (1993)	Analizaron la utilidad de la *Dissociative Experience Scale* (DES) como herramienta de detección temprana para desórdenes disociativos.
Griffiths, Myers y Talbot (1993)	Examinan la utilidad de la versión escalada del *General Health Questionnaire* (GHQ 28), como instrumento de *screening* para detectar casos psiquiátricos en esta población.
Jagger, Clarke y Anderson (1992)	Compararon el *Mini Mental State Examination* (MMSE) y el subtest *Information/Orientation* (IO) de la *Clifton Assessment Procedures for the Elderly* para detectar demencia en 1579 personas. Una submuestra de 438 sujetos se entrevistó usando el *Cambridge Mental Disorders of the Elderly Examination*.
Artículos anteriores a 1992	Uhlmann y Larson (1991), Wyshak, Barsky, Klerman (1991), Giles, Roffwarg, Rush y Guzick (1990), McCracken, Rubin y Poland (1990), Weinstein, Berwick, Goldman y Murphy *et al*. (1989), Murphy, Berwick, Weinstein, Borus *et al*. (1987) y Mari y Williams (1985).

Algunos de estos estudios se han llevado a cabo con poblaciones enteras de un país. Por ejemplo, Strand *et al*. (2003) comparan 4 instrumentos para evaluar la salud mental de la población noruega. Se utiliza la curva ROC para estimar los valores predictivos, y encuentran diferencias significativas entre las puntuaciones medias de hombres y mujeres. Se obtiene una AUC bastante buena, de 0.92.

Furukawa *et al*. (2003) utilizan dos nuevas escalas de detección para problemas psicológicos y las comparan con la encuesta nacional australiana de salud mental y bienestar, que utiliza una muestra representativa a nivel nacional. Se utilizan posteriormente las áreas bajo las curvas ROC para evaluar la capacidad predictiva de las tres escalas para detectar desórdenes de la DSM-IV y de la ansiedad.

Kessler *et al*. (2003) desarrollan un método para estimar la prevalencia de "enfermedad mental grave" (*serious mental illness*, o SMI). Se desarrollaron 3 escalas de detección para su posible uso en la encuesta nacional de hogares de la Administración de Servicios de Salud Mental y abuso de sustancias. Se utilizan las curvas ROC para la comparación de la precisión diagnóstica y se encuentran dos de ellos como más eficaces que los otros, en un estudio con una muestra de conveniencia de 155 sujetos.

Holi, Marttunen y Aalberg (2003) comparan dos versión del cuestionario de salud general y el *checklist* de síntomas como instrumentos de detección psiquiátrica en Finlandia, para lo cual utilizan las curvas ROC. Se obtienen resultados similares, y se sugieren puntos de corte óptimos para los instrumentos.

También queremos resaltar el interesante estudio de Freeman, Alegria, Vera, Muñoz *et al* (1992), que utilizaron el análisis ROC para la predicción del uso de servicios de salud. Estos autores examinaron la contribución de 4 dominios del Help Seeking Decision Making Model para predecir el uso de los servicios de salud mental por 1598 personas entre 18 y 64 años. La metodología combina el análisis de regresión logística con las curvas ROC. Se utilizaron las curvas ROC para comparar e interpretar la contribución relativa de un dominio de predisposición, un dominio de salud física y mental, un modelo restrictivo y un dominio organizacional para clasificar usuarios de no usuarios correctamente en los servicios de salud mental. Los descubrimientos sugieren

que la comparación de curvas ROC ayuda a describir e interpretar los dominios del modelo que son relevantes para hacer predicciones sobre quién y quién no usará los servicios mentales durante un periodo de 1 año.

Por último señalar el interesante trabajo de Castro (1999) sobre el juego patológico.

3.3.3 Predicción del reintento de suicidio

Muy recientemente se han aplicado estas técnicas para la detección del reintento de suicidio, que nos han parecido muy interesantes:

- Nimeus, Alsen y Traeskman (2002) evalúan items individuales y puntuaciones totales de la Escala de Intento de Suicidio (*Suicidal Intent Scale*, SIS) para evaluar su utilidad para predecir el suicidio. Se estudian 555 pacientes que fueron evaluados con esta herramienta poco después de un intento de suicidio. Después se les siguió un tiempo medio de 4.5 años, y en este periodo, 22 (4%) habían cometido suicidio. Se comprueba que los que así hicieron puntuaron más alto en el SIS, y para evaluar la relación formulan una curva ROC. Son capaces de obtener una sensibilidad del 90% y una especificidad del 60.3% a partir de puntuaciones de 19 en personas con más de 55 años.

- Osman *et al*. (2003) evaluaron la estructura factorial, fiabilidad y validez de un inventario de "ideación del suicidio" en una muestra de jóvenes de instituto (14-19 años). Se utiliza regresión logística y curvas ROC para identificar puntos de corte óptimos para las escalas positiva y negativa.

- Nimeus, Alsen y Traeskman-Bendz (2000) utilizaron una escala de evaluación de suicidio, conocida como SUAS, construida para medir el riesgo de cometer suicido en el tiempo. Su validez predictiva fue evaluada comparando las puntuaciones con las de otras escalas y se relacionaron con diagnósticos psiquiátricos y la edad, incluyendo otras enfermedades concordantes. Además de la edad avanzada, puntuaciones altas en SUAS fueron predictores significativos del suicido. Desde un punto de vista de la capacidad predictiva que proporciona la ROC, los autores

identifican las puntuaciones de corte que por ellas mismas y en combinación con otros factores demográficos son de importante valor en la evaluación del riesgo de suicidio después de un intento.

- Osman *et al*. (2001) destacan que el riesgo de suicidio lo predice la ideación y los intentos previos. Sin embargo estos autores señalan que se ha prestado poca atención al desarrollo de medidas validadas de conducta pasada de suicidio. Este estudio evaluó fiabilidad y validez de una medida breve de autoinforme de conducta pasada de suicidio con el cuestionario Suicidal Behaviors Questionnaire-Revised (SBQ-R). Los análisis de regresión logística proporcionaron apoyo empírico sobre la utilidad de este instrumento como una medida de riesgo de suicidio para distinguir entre los participantes en riesgo y sin riesgos.

Otro objetivo interesante de investigación para la detección es la anorexia. Al-Adawi *et al*. (2002) evaluaron la validez de una prueba de "actitud hacia la comida" para identificar la presencia y severidad de patologías relacionadas con la comida en hombres y mujeres de Omán, adolescentes urbanos, y para establecer puntos de corte que se ajustaran a los identificados con las "entrevistas de *gold standard*". Se utilizó una curva ROC para calcular el poder discriminativo para cada posible umbral, y se consiguen los siguientes resultados: un punto de corte de 10 dio el mejor compromiso entre sensibilidad y especificidad (64%) y especificidad (38%).

3.3.4 Detección de adicciones

Las adicciones es otro campo de aplicaciones clásico para el análisis de curvas ROC: Knight *et al*. (2003) comparan la validez de una escala de identificación de problemas con el alcohol, denominada AUDIT, con otras denominadas POSIT y CAGE. Estas pruebas se pasaron a pacientes de entre 14 y 18 años que llegaban a las unidades de cuidados rutinarios de un hospital, y se evaluaron también mediante la entrevista estándar de DSM-IV para diagnosticar abuso y dependencia del alcohol. Se utiliza ROC para encontrar puntos de corte óptimos.

Hinkin *et al*. (2001) examinaron la sensibilidad y especificidad de una versión modificada del CAGE, una medida de detección rápida utilizada para este propósito con personas mayores adictas. Se utilizó una revisión retrospectiva de las historias clínicas de 976 pacientes detectados por un programa de detección de individuos con problemas de adicción a alcohol y drogas. Los estudios de curvas ROC mostraron excelente sensibilidad pero poca especificidad, y omitiendo un ítem del CAGE mejoró significativamente la especificidad con un pequeño descenso de la sensibilidad, por lo que proponen esta versión modificada del CAGE como herramienta de detección temprana para adicción a alcohol o drogas en la población geriátrica.

Hay algunos estudios por autores españoles. Cuevas *et al*. (2000) evaluaron la validez de la Escala de Severidad de la Dependencia como prueba de detección rápida para la dependencia a la benzodiazepinas por parte de usuarios de las mismas. Los análisis ROC se utilizaron para determinar qué puntuación de corte permitió la mejor combinación de sensibilidad y especificidad, además de señalar un punto de corte. Se encuentra un área bajo la curva de 0.991.

Amador, Forns y Martorell (2001) estudiaron las propiedades psicométricas de las valoraciones de profesores y padres sobre el síndrome de déficit de atención por hiperactividad. Los resultados se evaluaron mediante el análisis de la curva ROC y otras pruebas estadísticas, que permiten establecer perfiles del síndrome y puntos de corte para las escalas de puntuación de padres y profesores.

Gual y otros (2002) realizan un estudio para identificar versiones cortas de la prueba *Alcohol Use Disorders Identification Test* (AUDIT) y para evaluar su eficacia como pruebas de detección rápida en contextos de cuidado primario. Se utilizó el diagnóstico de los clínicos como "gold standard" para evaluar la eficacia predictiva de tres formas del AUDIT. Se encuentra que hay dos formas cortas del AUDIT que son tan eficaces como la prueba completa para el propósito del estudio.

Kaye y Darke (2002) evaluaron la eficacia de un cuestionario de dependencia de la cocaína, denominado SDS, y para determinar la puntuación de corte que mejor discrimina entre la presencia o ausencia de un diagnóstico DSM-IV de la dependencia a la cocaína. Los participantes fueron 142 consumidores, y el rendimiento diagnóstico se

midió mediante el análisis ROC, que revelaron que este instrumento tiene una alta utilidad diagnóstica para la detección de la dependencia a la cocaína.

Gordon *et al*. (2001) encuentran que pueden tener una buena capacidad diagnóstica de problemas con el alcohol con sólo 3 preguntas. Evaluaron el AUDIT, las 3 primeras preguntas del AUDIT-C, la tercera pregunta del AUDIT-3 y preguntas sobre cantidad y frecuencia. Se registraron datos de un total de 13.348 pacientes. Las curvas ROC les permiten hacer comparaciones sobre la capacidad predictiva global de cada uno de estos instrumentos, y encuentran que para población general, y con el propósito de detección, una versión de sólo 3 preguntas del AUDIT identificaba los sujetos en riesgo tan bien como el AUDIT completo cuando esos bebedores se identificaban por el criterio cantidad-frecuencia.

3.3.5 Análisis coste-beneficio en contextos clínicos

El disponer de un esquema unificado para el análisis coste-beneficio es una de las mayores ventajas de las curvas ROC. Reuben *et al*. (2003) llevan a cabo un estudio para determinar los costes relativos de cuatro estrategias de identificación de riesgo y comparan su rendimiento en la predicción de uso de hospital por diferentes subgrupos de personas mayores basándose en edad, sexo y utilización previa del hospital. Utilizan una muestra de la "Población establecida para estudios epidemiológicos sobre la tercera edad". Suponiendo que las intervenciones basadas en la detección temprana proporcionaran un beneficio total de 1000 dólares por caso verdadero positivos, y 400 dólares por cada caso falso positivo, la estrategia secuencial fue ligeramente menos costosa que la de sólo auto-informe. Ambas de cualquier modo son considerablemente más baratas que la estrategia de hospitalización. Se utilizó la curva ROC para hacer estas comparaciones.

Un artículo interesante sobre las implicaciones metodológicas de las medidas de significación clínica, es el de Kraemer *et al*. (2003). Estos autores señalan en el resumen que aunque no hay una prueba formal de "significación clínica", se puede sugerir utilizar una de entre 3 tipos de medidas. Éstas incluyen medidas de fuerza de asociación entre variables, magnitud de la diferencia entre grupos de tratamiento y comparación, y

medidas de potencia del riesgo. En cuanto a éstas últimas revisan 5 medidas: la razón de *odds*, la razón de riesgo, la reducción relativa de riesgo, y el número necesario para tratar. Finalmente proporcionan un "relativamente nuevo tamaño del efecto", o AUC, que señalan "por razones históricas irrelevantes a la discusión actual significa área bajo la curva ROC", que integra muchos de los otros y está directamente relacionado con la significación clínica.

3.3.6 Predicción del maltrato

Queremos resaltar en este punto las muy interesantes aplicaciones del análisis ROC para la detección o diagnóstico de comportamientos violentos, como el maltrato, que no son tan frecuentes como quizá debieran ser, vista su gran utilidad en muchos ámbitos. Destaca en particular el trabajo de Rice, Harris (1995). Estos autores señalan que hasta muy recientemente ha habido poca evidencia de la capacidad de instrumentos clínicos o actuariales para predecir el comportamiento violento. Más allá, una variedad confusa de medidas se ha propuesta para la evaluación de la precisión de las predicciones. Este artículo demuestra que el análisis ROC tienen ventajas sobre otras medidas en tanto en cuanto son simultáneamente independientes de la tasa base para violencia en las poblaciones estudiadas y de la puntuación de corte elegida para clasificar o identificar los casos que se predicen violentos. Como ilustración del valor de este enfoque las tasas base de violencia se alteraron con el uso de seguimiento de 799 hombres previamente violentos, de 3.5, 6 y 10 años. Las tasas base para el seguimiento de 10 años fueron también alteradas cambiando la definición de recaída violenta y examinando un subgrupo de alto riesgo.

Este artículo también muestra hasta qué punto los métodos de ROC se pueden utilizar para comparar el rendimiento de diferentes instrumentos para la predicción de la violencia. En particular, se ilustra cómo las ROC facilitan las decisiones sobre si, para una tasa base particular, el uso de un instrumento de predicción tiene garantías.

Otro artículo interesantísimo en esta línea es el de O'Brien, John, Margolin y Erel (1994), quienes examinan el acuerdo entre esposos en relación con la agresión física entre ellos y si los hijos fueron testigos de la agresión, basándose en resultados de

encuesta. Para explorar la capacidad diagnóstica de un informe conjunto de los padres como indicador de la exposición de los niños a la agresión en el matrimonio, se utilizaron las ROC. Este análisis sugiere que estos informes son igualmente diagnósticos.

3.3.7 Otros estudios dentro de la Psicología Clínica

Critchfield (1993) por su parte estudió los sesgos y discriminabilidad de autoinformes sobre elecciones realizadas por personas. Sus resultados indican que los análisis basados en la TDS (para estas tareas de autoinformes) pueden mejorar la descripción de la correspondencia entre autoinformes y sus referentes y contribuir a la identificación de las fuentes ambientales de control sobre los autoinformes verbales.

Por último, una de las aplicaciones más peculiares que hemos encontrado es la de Gable, Reis y Downey (2003) sobre las discusiones de pareja. En su resumen explican que, en la vida diaria, los componentes de la pareja producen conductas a través de las cuales se influyen mutuamente. Para comprender cómo estos intercambios les afectan, la investigación previa ha estudiado la congruencia entre los informes de los que producen la conducta, y las percepciones de quienes la reciben. Diseñaron por tanto una estrategia basada en la TDS clásica que combina elementos de 3 enfoques en un estudio diario sobre los miembros de 58 parejas no casadas. Estas personas informaban diariamente tanto de sus conductas como de las percepciones de las de su compañero o compañera. Un ejemplo del tipo de análisis que llevan a cabo es el mostrado en la figura 3.1.

		Action "I told my husband that I loved him today "	
		I did it	I didn't do it
Detection "My wife told me that she loved me today "	She did it	Hit	False Alarm
	She didn't do it	Miss	Correct Rejection

Figura 3.1. Aplicación de los conceptos de la TDS a un contexto de terapia familiar(tomado de Gable, Reis y Downey, 2003).

3.3.8 Aplicaciones en Psicología Educativa

Además de las aplicaciones ya mencionadas para la detección del maltrato, se pueden mencionar varias aplicaciones en Psicología Educativa. McConville y Cornell (2003) parten de la hipótesis de que las actitudes agresivas predicen la conducta agresiva en estudiantes de grado medio. Para ello estudian las actitudes y su correlación con 4 criterios para conducta agresiva, autoinformes de estudiantes de agresiones de sus compañeros, nominaciones de compañeros y profesor y de los encargados de la disciplina en la escuela. Encuentran "tamaños de los efectos" entre 0.59 y 0.75, concluyendo que la evaluación de las actitudes de los estudiantes hacia la agresión proporciona información concurrente y predictiva sobre conductas agresivas en la escuela media.

Meisels *et al*. (2001) señalan que los juicios del profesor sobre el aprendizaje de sus alumnos son un elemento clave en la evaluación del rendimiento. Este estudio examina aspectos de la validez de los juicios del profesor que se basan en un instrumento de evaluación del rendimiento, incrustado en el curriculum, para determinar si se puede confiar en los juicios del profesor sobre el aprendizaje de los estudiantes. Las curvas ROC se utilizaron para comparar la precisión de este procedimiento para categorizar estudiantes en términos de sus resultados. Se muestra que el instrumento correlaciona bien con una batería tipificada, administrada individualmente, y que los datos obtenidos del mismo tienen utilidad significativa para discriminar con precisión entre niños que están en riesgo de los que no lo están.

También se debe mencionar el estudio de Niederer, Irwin, Irwin y Reilly (2003) para identificar a los niños con especial capacidad para las matemáticas en Nueva Zelanda.

3.4 El análisis ROC en Psicología Social y Forense

Uno de los campos de aplicación del análisis de curvas ROC en más rápida expansión es el de la Psicología Forense. Grann, Belfrage y Tengstreom (2000) exploraron la validez predictiva de dos instrumentos de evaluación de riesgo entre los delincuentes con algún trastorno mental en Suecia: el HCR-20 y el VRAG. Las puntuaciones

"actuariales" se obtuvieron retrospectivamente en dos poblaciones: un grupo de 358 delincuentes con trastornos de personalidad y otro con 202 personas violentas diagnosticadas con esquizofrenia. La precisión predictiva se evaluó con la ROC utilizando como variable criterio la recaída en el delito en 2 años desde la liberación. Ambas escalas rindieron mejor en el grupo de delincuentes con problemas de personalidad que en el de esquizofrénicos, y el H-10 funcionó mejor que el otro instrumento en ambos grupos. El estudio encontró que los datos históricos mantienen una validez predictiva robusta en poblaciones como las del estudio, mientras que los factores clínicos y de gestión del riesgo pueden ser de mayor importancia en poblaciones de delincuentes en las que grandes trastornos son prevalentes.

Tengstroem, Grann, Langstroem y Kullgren (2000) utilizaron el PCL-R para probar la hipótesis de que la psicopatía predice el recidivismo violento en una cohorte sujeta a investigación psiquiátrica forense, que consistía de 202 delincuentes violentos varones, entre 16 y 67 años con esquizofrenia. La psicopatía se evaluó con puntuaciones retrospectivas basadas en archivos. El tiempo medio de seguimiento después de la detención fue de 51 meses. El 22% de los delincuentes tenía una puntuación mayor o igual a 26, y la tasa base para recidiva violenta durante el seguimiento fue del 21%. Un análisis de supervivencia mostró que la psicopatía estaba fuertemente relacionada con la recaída violenta. La curva ROC varió entre 0.64 y 0.75.

Sjoestedt y Langstroem (2002) exploraron la capacidad predictiva de 4 medidas de evaluación del riesgo entre violadores. Se siguieron 51 varones convictos de violación y diagnosticados con trastornos de personalidad en la evaluación psiquiátrica forense antes del juicio, en Suecia entre los años 1988 y 1990, y fueron seguidos con respecto a recaídas (detenciones) durante una media de 92 meses después de su salida de prisión o tratamiento psiquiátrico. Las tasas base para reconvicciones fueron de un 20% para delitos sexuales, 25% delitos con violencia no sexuales, y de cualquier tipo (incluyendo sexuales) del 39%. Sólo uno de los instrumentos proporciona precisión predictiva significativamente mejor que por azar, con un área bajo la curva de 0.73.

Sjoestedt y Grann (2002) señalan que hay un debate entre los enfoques actuarial y clínico para evaluar el riesgo de recidivismo violento, puesto que hay estudios previos sobre la validez predictiva que dicen que favorecen las evaluaciones actuariales. Sin

embargo, estos estudios han usado un resultado dicotómico como la variable criterio. Para no sólo predecir, sino también gestionar el riesgo de recidivismo y prevenir la violencia futura, hay una necesidad de trabajar en la variable criterio en las evaluaciones actuariales y también considerar la inminencia, frequencia, naturaleza y severidad del potencial caso en riesgo. Se revisó una base de datos de evaluaciones actuariales de riesgo que se referían a una cohorte de 5 años de todos los hombres adultos liberados de prisión después de haber realizado condena. Las conductas de recaída en delitos sexuales se codificaron para individuos que fueron condenados durante un periodo de seguimiento de 6 años. Los procedimientos actuariales tenían áreas bajo la curva ROC de 0.73 y 0.75 en relación con esta repetición de delito. Sin embargo, cuando los procedimientos se re-evaluaban utilizando varios resultados, la validez predictiva variaba entre 0.40 y 0.94 (área bajo las curvas). Los esquemas de evaluación del riesgo actuariales funcionaron bien para la recaída inminente y menos severa que para otros tipos de repetición de delito.

Hayes (2002) propone que el sistema legal realice identificaciones tempranas de individuos con discapacidades intelectuales. Se comprueba que las personas con una discapacidad intelectual están significativamente sobre-representados en el sistema de justicia criminal en muchas jurisdicciones occidentales. Para esta identificación temprana se comparan 3 pruebas y se obtienen datos de 567 adultos y jóvenes delincuentes. Se analizan las curvas ROC y se comprueba que son muy similares en la capacidad predictiva.

Kroner y Mills (2001) estudiaron la capacidad predictiva de varias pruebas con el propósito de detectar o evaluar el riesgo de conducta antisocial y nuevas detenciones.

Furlong, Bates y Smith (2001) llevan a cabo un estudio para predecir la posesión de armas en la escuela, mediante un análisis secundario de la encuesta de comportamiento de riesgo en la juventud. Señalan en su artículo que los psicólogos escolares son solicitados para ayudar a evaluar el nivel de riesgo por estudiantes específicos, y para ello necesitan considerar la propiedad técnica de cualquier prueba o procedimiento que se proponga para "predecir" la conducta violenta futura. En su estudio se examinaron las respuestas de 40.435 estudiantes de entre 9º y 12º grado de las encuestas de conducta de riesgo en la juventud de 1993, 1995 y 1997. Las conductas de riesgo auto-informadas

por los estudiantes y la experiencias se utilizaron para predecir la posesión reciente de armas en los campus escolares y para ilustrar el uso de las curvas ROC para evaluar el rendimiento de un test con propósitos predictivos. Un índice de 9 conductas de riesgo estuvo moderadamente correlacionado con la posesión de armas. Sin embargo, cuando se evalúa la precisión de utilizar este riesgo para predecir la posesión de armas, se encontró que el área bajo la curva ROC fue de aproximadamente 0.75. Además se identificó que había portadores de armas hasta nueve veces más frecuentes con puntuaciones de riesgo 0 que con puntuaciones de 7-9, y este resultado sugiere que la atención hacia los perfiles de riesgo de violencia puede estar promoviendo la falta de atención a otro grupo de estudiantes que pueden engancharse en conductas de alto riesgo en la escuela.

Muy interesante resulta la aplicación de J. Copas (1999) al campo de estimadores de riesgo y en particular a la predicción de la violación de la libertad condicional. En la figura 3.2 se muestran dos curvas ROC de la puntuación "*Parole Board*" (Comité de Libertad Condicional). Esta puntuación (Copas *et al*., 1996) es una función lineal de covariantes que incluye la edad, el número de condenas previas, el número de sentencias de custodia previas en la juventud y edad adulta, y una clasificación del crimen principal en seis categorías extensas de crimen. La curva ROC se estima a partir de E, que se define como "recidivismo" en los 2 años siguientes: E sucede si, en los dos años posteriores a abandonar la prisión, el condenado comete otro delito para el que se le condena. En este caso encuentran un área bajo la curva de 0.772. Este periodo de seguimiento de 2 años es por supuesto arbitrario. Una definición más estricta de recidivismo sería considerar el riesgo de volver a cometer un delito en 3 meses. Esta definición permitiría calcular otra curva, que se muestra por debajo de la anterior, para la cual el área es de 0.725. La diferencia sería muy pequeña. Las curvas ROC tal y como las muestra Copas en su artículo aparecen en la figura 3.1.

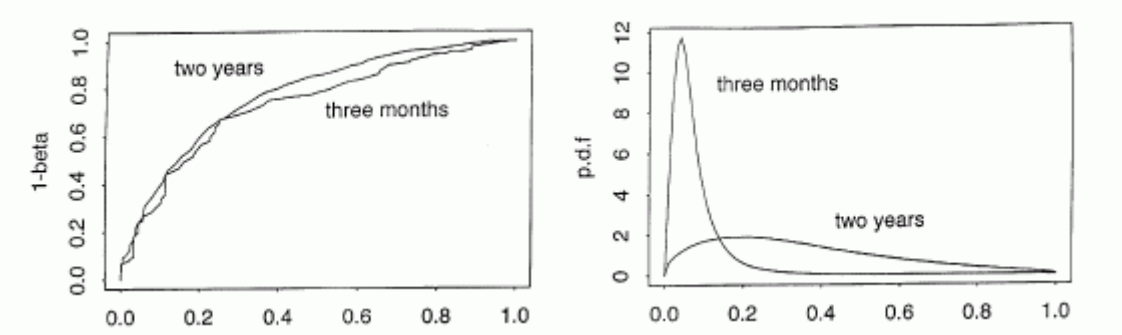

Figura 3.2. Curvas ROC para la predicción de la violación de la libertad condicional (tomado de Copas, 1999)

Existen más estudios para la detección temprana de problemas psiquiátricos en el campo legal. Por ejemplo, Tengstrom, Grann, Langstrom y Kullgren (2000). Estos autores señalan haber utilizado el software de análisis *MedCalc for Windows* 4.2.

La Psicología Forense ha encontrado muchas aplicaciones de las curvas ROC en los delitos contra la libertad sexual. Craig, Browne y Stringer (2003) señalan la gran demanda de evaluación precisas de los niveles de riesgo de esta población porque las decisiones sobre estas personas pueden tener un gran impacto sobre el acusado y la población general. Este artículo revisa los factores predictivos del recidivismo de la conducta de agresión sexual y cuestiona la validez de algunos de estos instrumentos por no capturar la naturaleza dinámica de estas conductas.

Harris *et al*. (2003) desarrollan un estudio en múltiples sitios sobre instrumentos actuariales de riesgo para delincuentes contra la libertad sexual. En particular, evalúan 4 pruebas en 4 grupos de delincuentes, que suman un total de 396 casos. Aunque los 4 instrumentos predecían de forma significativa, las áreas bajo las curvas ROC respectivas fueron consistentemente mayores para dos de ellos.

Edens *et al*. (2002) comparan la utilidad de dos medidas de rasgos psicopáticos para explicar la conducta antisocial grave entre delincuentes contra la libertad sexual encarcelados. Se utilizaron las infracciones disciplinarias de 92 de estas personas, y se sometieron a análisis ROC dos instrumentos para predecir el resultado, sin encontrar diferencias significativas entre ellas.

Uno de los temas de la Psicología Social aplicados al campo legal que se ha beneficiado del uso de las técnicas de análisis de curvas ROC son la evaluación de testimonios o de expertos. Por ejemplo, Lewis (1992) señala que el modelo de detección de señal, con dos gausianas, con criterio variable, proporciona una buena descripción del comportamiento del experto, y forma la base para la distinción entre la precisión y el rendimiento del experto. La precisión se define finalmente en términos de la ROC y es independiente de los criterios de decisión y probabilidades a-priori del evento. El rendimiento del experto no sólo depende de su precisión, sino también de los criterios de decisión utilizados por el experto y las probabilidades a-priori del evento.

Otro artículos que aparecen en la literatura es el de Schum (1981), quien realizó un estudio de los factores que actúan en conjunto para determinar el valor inferencial de evidencia testimonial directa y circunstancial, para lo que utiliza la TDS y la teoría de inferencia bayesiana. La TDS permite estudiar, sin confundirlos, los factores que influyen en la sensibilidad del observador y factores como las expectativas del observador, motivación y objetivos que binfluyen en la decisión. Las tasas de éxito y falsos positivos, son ingredientes de credibilidad para las formulaciones que luego se utilizarán en la teoría bayesiana sobre el valor inferencial de un testimonio. Kushnir y Duncan (1978) estudian fenómenos de facilitación social desde la perspectiva de la TDS. Este enfoque permite la distinción entre los sesgos del estímulos (p.ej., efectos en la sensibilidad) y tendencias de respuesta (p.ej., efectos en el criterio).

3.5 Aplicaciones del análisis ROC en tecnologías de la información

En la actualidad existen muchos estudios a medio camino entre la bioquímica más tradicional y la computación. Como muestra de estos estudios, señalaremos el de Toivonen, Srinivasan, King, Kramer y Helma (2003). Dentro de lo que se denomina "retos" (*challenge*), el objetivo es desarrollar modelos "*in silico*" (computacionales) para predecir la carcinogénesis química en cánceres causados por factores ambientales. Estos autores utilizaron 14 grupos de máquinas de aprendizaje que generaron 111 modelos, y usaron lo que ellos denominan el espacio ROC para que los modelos fueran uniformemente comparables independientemente de la función de coste. Desarrollaron

un modelo estadístico para probar si un modelo rinde significativamente mejor que por azar en el espacio ROC, en el que prueban todos los modelos y encuentran dos ganadores. Un ejemplo del análisis de curvas ROC realizado aparece en la figura 3.3.

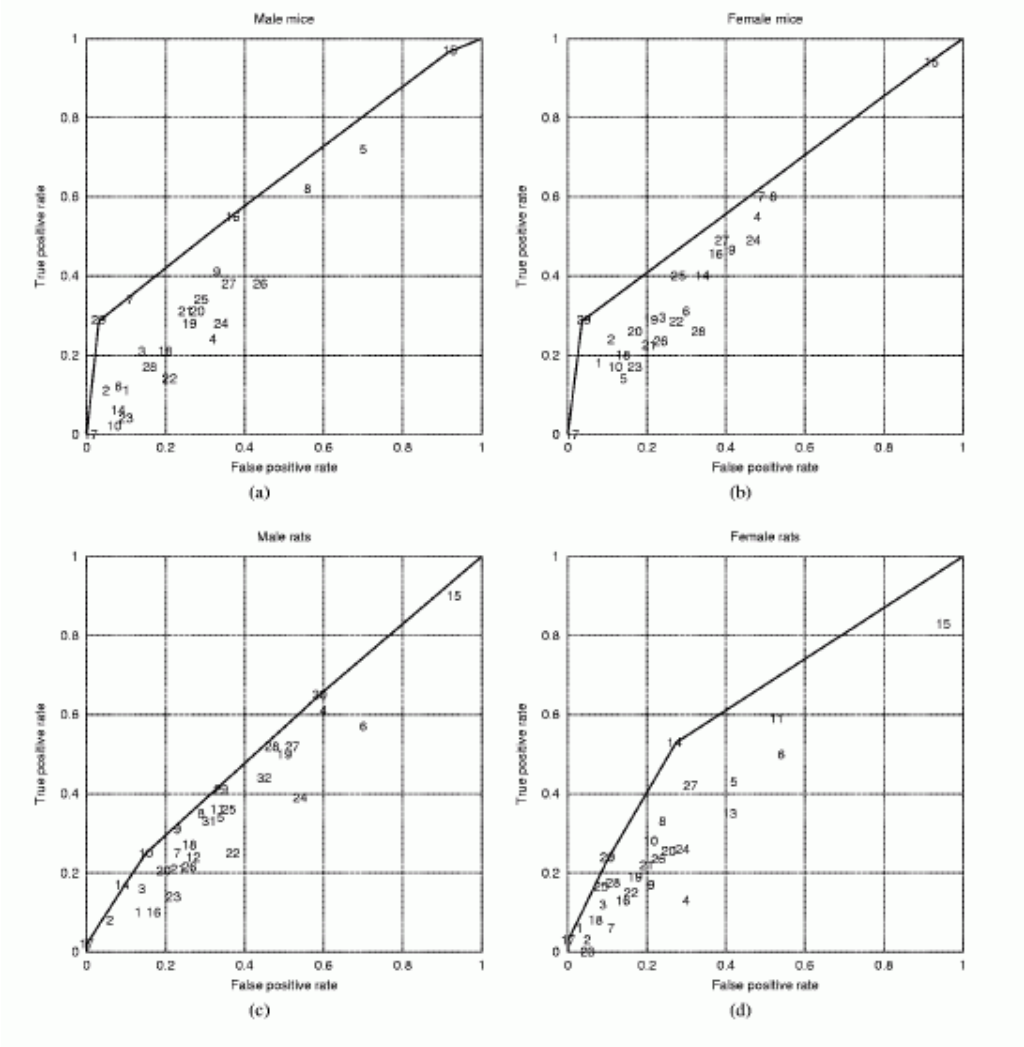

Figura 3.3. Aplicaciones de las curvas ROC para evaluar clasificadores, tomado de Toivonen, Srinivasan, King, Kramer y Helma (2003).

Por otro lado, resulta interesante ver que el análisis de curvas ROC se ha aplicado con mucho éxito a las tecnologías de la información, o informática. Hay artículos pioneros, como el de Koenemann y Belkin (1996), o los propios de Swets (1969). Se trata en muchos casos de evaluar la eficacia de algoritmos de clasificación o de procedimientos o procesos completos de recuperación de información.

Por ejemplo, Kwong Bor y Kantor (2000) evaluaron dos variables predictivas para la efectividad de una fusión de datos, y comprueban la efectividad de cada modelo de fusión mediante el análisis de curvas ROC.

Una de las aplicaciones más interesantes en áreas ajenas a la Psicología es el tema de los seguros. Aquí destaca el completísimo artículo de Viaene, Derrig, Baesens y Dedene (2002) sobre el estado del arte sobre las técnicas de clasificación para la detección de fraude en los partes de seguros de automóviles. Los autores primero y tercero, junto con otros 4 autores más, aplican un enfoque similar para el "*scoring*" de crédito (otorgar una puntuación basada en el riesgo de no devolución del préstamo) en Baesens *et al*. (2002). En este estudio utilizan datos de entidades financieras del Benelux y del Reino Unido, y comparan diferentes tipos de clasificadores:

- Los más conocidos, regresión logística, análisis discriminante, vecino k-más próximo, redes neuronales
- Los más avanzados, menos conocidos, como los algoritmos basados en kernels avanzados, tales como las "*support vector machines*" y "*least-squares support vector machines* (LS-SVMs)".

La evaluación del rendimiento se evalúa utilizando la precisión de clasificación y el área bajo la ROC. Estos autores encontraron que los métodos LS-SVM y las redes neuronales proporcionan muy buen rendimiento, pero también clasificadores más simples como la regresión logística y el análisis discriminante lineal funcionan muy bien para el "*scoring*" de crédito.

Otra aplicación peculiar, aunque muy interesante, es la de Graham, Russel, Stevenson y Torbey (2001), sobre la predicción de resultados a los accionistas después de una bancarrota. Estos autores usan el análisis de curvas ROC para comparar dos modelos predictivos, uno basado en datos exclusivos de la firma solicitando la quiebra, y el otro mixto, con datos de mercado y propios de la firma. Ambos modelos predicen aproximadamente igual.

Aunque no directamente relacionado con las tecnologías de información, reseñamos aquí como prácticamente único representante de las posibles aplicaciones al campo del control de calidad, el artículo de Baker (1975), quien aplica la TDS para la evaluación

de eficacia de inspección de calidad de control industrial. Discute la similaridad de modelo TDS con la teoría de muestreo y aceptación de calidad en control de calidad. En su artículo se muestran métodos de calcular medidas TDS de detección de fallos por inspectores y estándares de calidad subjetiva. Se ilustran ejemplos para ilustrar su utilidad para diagnosticar las causas de rendimiento insatisfactorio de inspecciones.

3.6 El análisis ROC en la investigación de mercados

Por el desarrollo lógico de esta tesis, este apartado debiera estar el primero, o quizá el segundo o el tercero, en este capítulo. Pero es que una búsqueda bibliográfica en las revistas de investigación de mercados, o "Marketing" en general, llevada a cabo sólo unos pocos meses antes de la terminación de este trabajo, mostró sólo unos pocos resultados en los que apareciera ROC (o "*Receiver Operating Characteristic*", que es el término que aparece normalmente en las palabras clave). Se trata, por tanto, de una técnica que todavía ha encontrado poca o prácticamente ninguna difusión en este campo. Con una única excepción que será mostrada al final de este apartado, no porque sea la menos representativa, sino porque es la mejor, y permite apreciar un salto cualitativo importante en la aplicación de las curvas ROC en lo que ha venido en llamarse "minería de datos". Se trata del excelente reciente trabajo de Baesens et. al (2004) sobre la aplicación de redes bayesianas para la predicción de la tendencia de gasto de clientes.

Señalaremos brevemente algunos de los restantes artículos, que tienen que ver con esta tesis más por omisión (puesto que utilizan procedimentos de comparación de modelos mucho menos potentes y válidos que curvas ROC) que por presencia de técnicas de utilidad para nuestro propósito.

Una de ellas es la de Thieme, Song y Calantone (2000). Thieme *et al*. (2000) desarrollan un sistema de soporte a la decisión mediante redes neuronales artificiales y demuestran cómo puede guiar a gestores cuando se trata de tomar decisiones complejas sobre nuevos desarrollos de productos. No utiliza ROC y es un ejemplo de las limitaciones de las medidas de eficacia cuando no se utilizan ROCs.

Estos autores utilizan datos de 612 proyectos para comparar un método basado en redes neuronales con los métodos tradicionales para predecir el éxito de nuevos proyectos de productos. Una vez aplicados los modelos bajo prueba, se mide su eficacia en términos de la Suma de Errores cuadráticos (SSE) y Media del Error Cuadrático (MSE). Con ese criterio obtienen la conclusión de que el método de redes neuronales mejora a los siguientes métodos: vecino k-más próximo (*k-nearest*), regresión logística, regresión mínimos cuadrados y análisis discriminante. Posteriormente dicotomizan a partir de 0.5 y calculan las matrices de confusión 2 x 2. Aunque con este enfoque capturan la complejidad de las decisiones a partir de esta tabla, no continúan su análisis ni con los conceptos de sensibilidad o especificidad, ni con el cálculo de las curvas ROC.

El siguiente estudio lo mencionamos por curiosidad, no por un interés directo en el tema. Se trata del estudio de Platter *et al*. (2003) que estudia la relación entre valoraciones de algunos parámetros sensoriales emitidos por consumidores, y su relación para la aceptación de filetes de carne. En particular evalúan la relación de una puntuación de "*marbling*", que mostró una débil relación r^2 de 0.053, aunque significativa, con la aceptación de dichos filetes.

Pero desafortunadamente sólo muestran el porcentaje de clasificación correcta de un modelo logístico según se muestra en figura 3.4., que es lo que luego veremos como "gráfico de elevación" o *lift-chart*. Esta herramienta es muy común en el campo de Marketing y veremos en un apartado específico sus diferencias, y sus limitaciones, con el enfoque de curvas ROC.

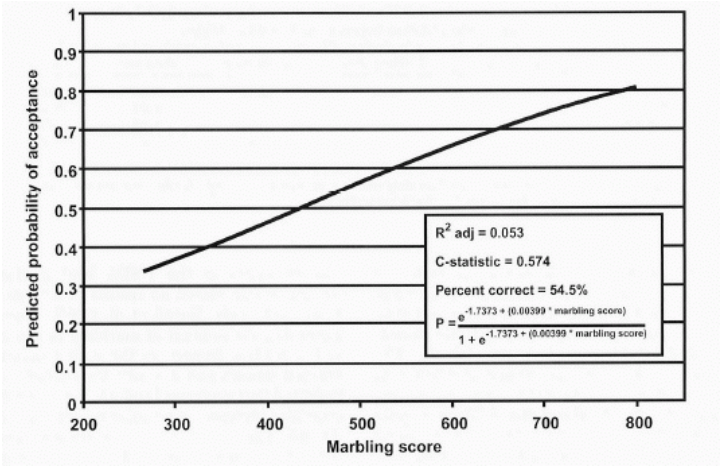

Figura 3.4. Gráfico de elevación de la puntuación "marbling" para predecir la probabilidad de aceptación de la carne. Tomado de Platter et al. *(2003)*

Otro artículo similar es el de Brethour (2000). En este estudio se señala la potencia del análisis de curvas ROC para determinar puntos críticos operativos para seleccionar grupos con porcentajes de calidad determinados. El procedimiento de medida que se pretendía evaluar era por ultrasonidos en una cabaña de ganado. Cuando se puede establecer un estimador a priori de la prevalencia de diferentes grados de calidad, los procedimientos de ROC permiten incorporar tasas de error para estimar el porcentaje de la cabaña de ganado que serán seleccionados en un grupo y otro. Además, ofrece la capacidad de llevar a cabo un análisis de coste-beneficio. El resultado se muestra mediante curvas ROC (figura 3.5).

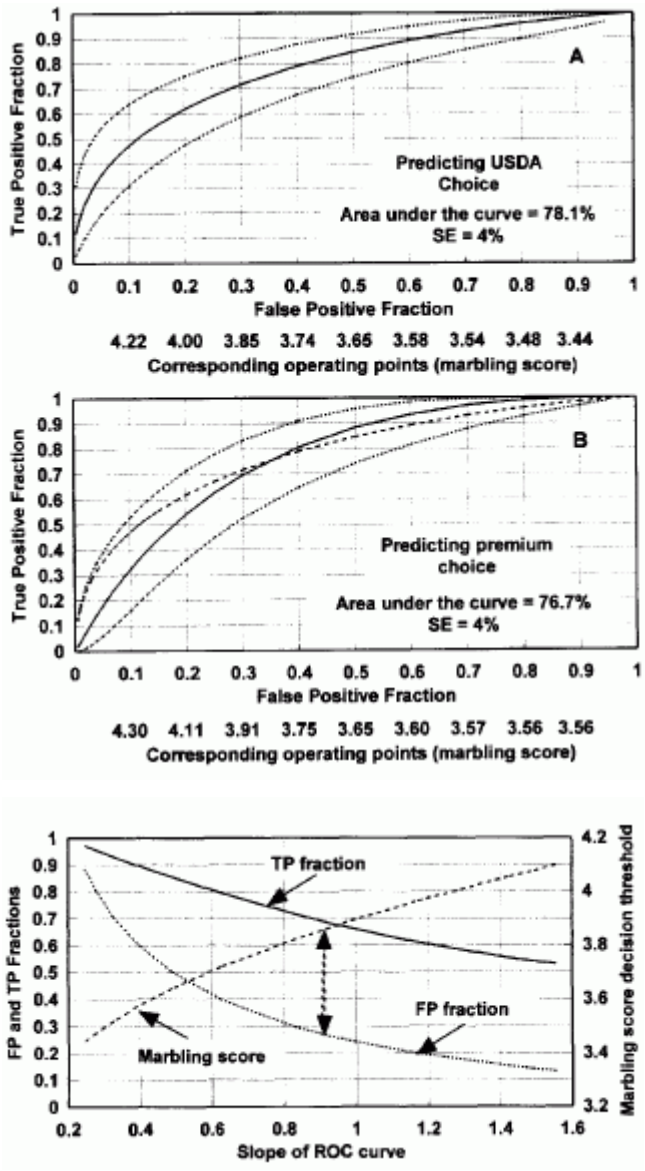

Figura 3.5. Aplicación de las curvas ROC en evaluación de la calidad de la carne (tomado de Brethour, 2000).

3.6.1 Redes bayesianas para la modelización de la fidelidad y curvas ROC

Uno de los pocos artículos dentro del campo de investigación de mercados con importante contribución del análisis ROC es el de Baesens *et al*. (2004). En su estudio se concentran en introducir una medida de la evolución futura de gasto de un cliente que permita mejorar la toma de decisiones por parte del departamente de Marketing relacional. Desde el punto de vista de Marketing, el objetivo es predecir si un cliente recientemente adquirido incrementará o no su gasto a partir de la información de la compra inicial. Esta tarea es esencialmente de clasificación binaria, y la principal novedad de su estudio reside en la comparación y evaluación de varios clasificadores basados en redes bayesianas que, mediante técnicas estadísticas y otras de inteligencia artificial, permiten realizar esta tarea.

En su introducción estos autores señalan que no se encuentran estudios publicados que hayan intentado predecir la variable resultado como incremento o decremento de gasto futuro, pero que existen múltiples estudios sobre el tema de la fidelidad. El objetivo será por tanto poner a punto y comparar técnicas de clasificación desarrolladas recientemente para la clasificación óptima de los clientes en dos grupos relevantes, esto es, clientes con gasto decreciente vs. clientes con gasto creciente. Las técnicas utilizadas tienen que proporcionar buenos resultados de predicción en la clasificación y también representar las relaciones de independencia marginal y condicional entre las variables y cómo conjuntamente afectan a la decisión de clasificación.

Como punto de partida utilizan un clasificador "*naive*" de Bayes, al que gradualmente se le van eliminando restricciones sobre la estructura de la red, y se investigan a continuación los "*Tree Augmented Naive Bayes classifiers* (TANs)", seguidos por los clasificadores de red bayesiana completamente sin restricciones. Todos los clasificadores se evalúan en su capacidad predictiva o de clasificación mediante el área bajo la curva ROC, que estos autores denominan AUROC. Estos autores señalan que esta técnica de evaluación de la capacidad predictiva efectivamente separa el rendimiento en la clasificación de estos factores. Otro criterio de elección de modelos es la parsimonia o economía de parámetros de los mismos.

Uno de los principales problemas en el Marketing relacional actual es el de la antiguedad del cliente como criterio de fidelidad. Existen resultados recientes en varios sectores económicos que no han encontrado evidencia que sugiera que los clientes con mayor antiuguedad y con una conducta estable de compras sean necesariamente más económicos para servir, menores sensibles a los precios, o más efectivos para atraer más negocio a la compañía. Algunos autores que mencionan Baesens *et al*. (2004) mostraron que las contribuciones de este grupo podían estar generalmente en descenso, y que los mayores beneficios podían venir de clientes con vida corta pero con altos beneficios. Hay que destacar que se trataba de una compañía de venta por correo, así que estos resultados pueden estar circunscritos a este sector económico en particular.

En estos estudios a los que se hace referencia se ilustran los errores asociados a utilizar una gran parte del presupuesto de Marketing en clientes que han sido buenos clientes en el pasado frente a los que lo son por un periodo corto de tiempo, los que han sido llamados "mariposas" (*butterflies* en su original inglés). En el sector de venta por correo se ha encontrado que la conducta de recompra puede ser modelizada efectivamente mediante una combinación (a menudo lineal) de variables conocidas como RFM, esto es, la Recencia de la última compra de un clinete, la Frecuencia media de las compras de ese cliente, y el valor Monetario medio gastado en las ocasiones de compra del cliente. De aquí que el grupo de clientes denominados "mariposas", dado que tienen un valor monetario histórico alto, tiendan a estar sobrerrepresentados en las campañas de "*mailing*". Una estimación de la pendiente futura del ciclo de vida del cliente, esto es, la evolución del gasto del cliente, podría entonces proporcionar la comprensión necesaria para el proceso de toma de decisiones y la comprensión de la relación entre la pendiente y otras variables, tales como el gasto del ciente, que podría generar información cualitativa rica para los decisores del departamento de Marketing. Por ejemplo, se podría concentrar la capacidad de Marketing directo de inversiones a largo plazo a inversiones en promociones en los que se puede conseguir un retorno a corto plazo, o se podría sencillamente decidir no concentrarse sobre estos clientes. En cualquier caso, el conocimiento a priori de esta pendiente del ciclo de vida del cliente sería una información enormemente útil. Para cumplir sus objetivos, los autores deciden utilizar un criterio dicotómico, representado por la siguiente pregunta: "¿incrementarán o decrementarán los nuevos clientes su gasto después de sus primeras experiencias de compra?".

En la literatura de Marketing, estos problemas de clasificación binaria se han afrontado utilizando métodos estadísticos tradicionales, como el análisis discriminante y la regresión logística, mediante modelos estadísticos no paramétricos, como el vecino k-mas próximo, y los árboles de decisión, y por redes neuronales. La novedad es que estos autores utilizan clasificadores de redes bayesianas, que han comenzado a aparecer recientemente en la literatura de inteligencia artificial.

Los clasificadores de redes bayesianas son modelos probabilísticos de "caja blanca" que facilitan una comprensión clara de las dependencias que subuyacen al dominio que se estudia.

Una red bayesiana es básicamente un modelo estadístico que hace posible calcular la distribución de probabilidad conjunta a posteriori de cualquier subconjunto de variables estocásticas no observadas, si las variables en un subconjunto complementario son observadas. Esta funcionalidad hace posible usar una red bayesiana como clasificador estadístico utilizando la regla "el ganador lo lleva todo" a la distribución de probabilidad de la distribución de probabilidades a posteriori para los nodos no observados. El supuesto que subyace a la regla de "el ganador lo lleva todo" es que todas las ganacias y pérdidas son iguales. No entraremos aquí en la descripción de las redes bayesianas por su complejidad, sólo señalar que la revisión que presentan estos autores del estado del arte de estas técnicas es de lo mejor que se puede leer sobre el tema. Sí señalaremos para tener suficientemente completa esta revisión los tipos de redes que contemplan:

- Clasificador "*naive*" de Bayes.
- Clasificadores "*Tree Augmented Naive Bayes*" (TANs)
- Clasificadores generales de red bayesiana (GBN)
- Clasificadores multired de red bayesiana.

La investigación empírica se desarrolló sobre los datos de compra en el punto de venta ("*scanner data*") de un vendedor del sector conocido como DIY (*Do-It-Yourself*) en Bélgica. Los datos se obtuvieron mediante las tarjetas de fidelización, que llevan en uso desde 1995. Se utilizaron 4 años de información. Se utilizaron los dos primeros años de información para comprobar que los clientes en la muestra de estudio eran nuevos clientes. Finalmente se estudiaron 3827 clientes, en 15 atributos. Para estudiar la calidad

de los modelos, se dividió la muestra aleatoriamente en dos partes. Dos terceras partes se usaron para el aprendizaje de los clasificadores, mientras que el tercio restante se usó para estimar la conducta de generalización de los clasificadores.

La pendiente de cada cliente se calculó mediante un modelo de regresión lineal sobre las contribuciones históricas de cada cliente. A continuación se discretizó esta pendiente en positivos y negativos, en función de si la pendiente era positiva, y por tanto la tendencia de gasto creciente o negativa, o la tendencia de gasto era decreciente. Estos autores validaron algunas de las conclusiones de estudios anteriores en el sentido de que la pendiente de los clientes con mayor antiguedad era generalmente decreciente, dado que sólo el 28% de esos clientes en la base de datos mostraron una pendiente negativa.

Las variables independientes se dividieron en cuatro grupos:

- Uno para medir el volumen de las compras durante los primeros seis meses como cliente. Estas variables se pueden considerar como la "profundidad" de las compras de los clientes.
- El segundo grupo contiene las variables que miden la "anchura" de las compras, o proporción de categorías en las que se producen las compras.
- Un tercer grupo de variable capturó la tendencia a la negociación y la sensibilidad a los precios del consumidor.
- Finalmente se introducen 3 medidas para evaluar las evoluciones dentro de los seis primeros meses.

El rendimiento de los clasificadores se cuantificó mediante la precisión de la clasificación que mide el área bajo la curva ROC (AUROC) en términos de estos autores. La precisión de la clasificación es sin duda la medida más normalmente usada de rendimiento de un clasificador. Estos autores señalan que frente a la medida simple del total de casos verdaderos predichos, la curva ROC permite estudiar la conducta de un clasificador sin depender del punto de corte, o del coste de la distribución, y proporciona una medida única del rendimiento del clasificador construido. Una interpretación intuitiva del AUROC es que proporciona un estimador de la probabilidad de que una instancia elegida aleatoriamente de la clase positiva será evaluada u ordenada más alto que una instancia seleccionada aleatoriamente de la clase negativa.

Estos autores citan además los trabajos sobre contraste estadístico del área bajo las curvas ROC de Hanley y McNeil (1982, 1983) y De Long *et al*. (1989), que serán los que utilicen en las comparaciones. Las áreas bajo las curvas respectivas oscilaban entre el 0.723 y 0.75. Utilizando el enfoque anterior de contraste estadístico no paramétrico de las áreas bajo las curvas, muestran que los mejores clasificadores son el "naive" y el GBN, y que no hay diferencia significativa en el rendimiento de los clasificadores GBN (clasificador general bayesiano) y el clasificador bayesiano "*naive*", el más simple, con un nivel de significación del 5%.

Además de evaluar el rendimiento en la clasificación también se investigó la complejidad de los modelos generados, porque es obvio que será preferible un modelo más sencillo. El modelo bayesiano "*naive*" tenía 16 nodos y 15 arcos, mientras que el GBN tuvo 4 nodos y 6 arcos. Por este motivo, además de por el anterior, éste modelo es el que se puso en práctica. Utiliza sólo 3 variables compiladas de los registros de compra de los primeros seis meses en el ciclo de vida del cliente, y por lo menos para el sector en el que se hizo el estudio, se puede predecir el signo de la pendiente de este ciclo con una precisión del 75%. Las variables en concreto son la contribución total del cliente, el número total de artículos comprados y el porcentaje máximo de productos comprados en una familia de productos. Las primeras variables son representativas de la "profundidad" de la compra, y la última de la "anchura" de la compra.

Esto implica que los clientes que tienden a incrementar su gasto a lo largo de su vida con la empresa inicialmente gasta menos dinero en un menor número de artículos, comprando de un menor conjunto de categorías de productos. Alternativamente, los clientes que gastan mucho dinero inicialmente en muchos artículos y que compran productos a lo largo de muchas categorías diferentes tienen a reducir su gasto en el futuro. Baesens *et al*. (2004) concluyen diciendo que esta información puede probarse muy valiosa para una empresa como la de este caso como punto de partida para investigar por qué los clientes con alto gasto generalmente decrecen su gasto a lo largo del tiempo.

De forma muy interesante, Baesens *et al*. (2004) muestran que estos clasificadores de redes bayesianas son una herramienta muy poderosa, incluso desde el punto de vista conceptual, y presentan una propuesta para clasificación básica, o a priori, de la base de

datos de cliente en función de dos tablas de clasificación cruzada, una sobre la antiguedad y el signo de la pendiente de compra, y otra a partir del importe de la primera compra y el signo de esta pendiente. Estas tablas se muestran en la figura 3.6.

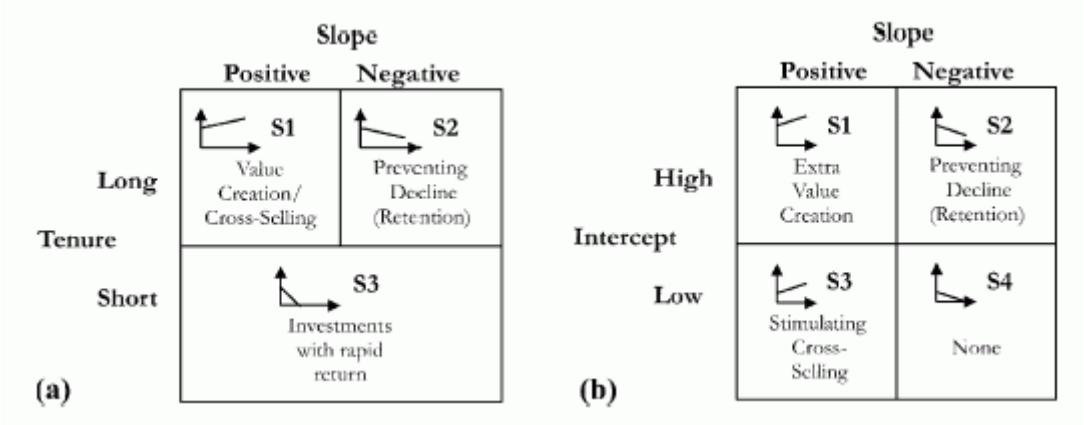

Figura 3.6. Posibles esquemas de clasificación de clientes a priori, propuesto y tomado de Baesens et al. (2004)