

UNIVERSIDAD DE ALMERÍA



Departamento de Estadística y Matemática
Aplicada

MODELIZACIÓN Y PREDICCIÓN

EN

ESTADÍSTICA UNIVERSITARIA

TESIS DOCTORAL

María Morales Giraldo

Almería, Mayo de 2006



**MODELIZACIÓN Y PREDICCIÓN
EN
ESTADÍSTICA UNIVERSITARIA**

TESIS DOCTORAL

María Morales Giraldo

DIRECTORES:

Carmelo Rodríguez Torreblanca

Antonio Salmerón Cerdán

Mayo 2006

La memoria titulada

Modelización y predicción en estadística universitaria

que presenta Dña. María Morales Giraldo, para optar al grado de **Doctor por la Universidad de Almería**, ha sido realizada en el *Departamento de Estadística y Matemática Aplicada* de la *Universidad de Almería*, bajo la dirección del Dr. D. Carmelo Rodríguez Torreblanca y el Dr. D. Antonio Salmerón Cerdán.

Almería, Mayo de 2006.

Fdo: Dña. María Morales Giraldo.

Fdo: Dr. D. Carmelo Rodríguez Torreblanca.

Dr. D. Antonio Salmerón Cerdán.

Dedicado a Antonio, a mis padres y hermano

Agradecimientos.

Es difícil resumir en poco espacio mi más sincero agradecimiento a tantas personas que me han ayudado y han estado a mi lado en estos años.

En primer lugar, quisiera agradecer a mis padres su continuo apoyo, su confianza y mi educación tanto académica como ética, y por infundirme los principios que más valoro. Pero ante todo, gracias por su regalo más preciado, mi hermano José, que siempre está junto a mí.

A mi marido Antonio, que al mirar hacia atrás, siempre ha estado a mi lado: gracias por su aliento durante mis estudios, a su apoyo como compañero en el trabajo, gracias por su inestimable labor como director de tesis, al tiempo y al esfuerzo dedicado generosamente a ella. A su comprensión y ánimos en los momentos más duros. Gracias, en definitiva, por cada día, y por mostrarme que no hay mayor libertad que la de vivir preso en alguien.

También quisiera tener un recuerdo especial hacia mi profesor del instituto, José Moya, quien me descubrió la belleza de las Matemáticas y por el cual elegí el camino que hoy culmina en este trabajo.

En el mundo académico quisiera dar las gracias a Carmelo Rodríguez, mi otro director de tesis, por el tiempo dedicado a la supervisión de este trabajo, pero ante todo, gracias por su confianza, amistad y por darme la oportunidad de trabajar en la Unidad de Coordinación de Datos y Estadística Universitaria de la Universidad de Almería, que ha supuesto una de mis mayores satisfacciones en estos últimos seis años, teniendo el privilegio de trabajar con dos ejemplos a seguir, desde un punto de vista tanto profesional como personal: la directora de la Unidad, M^a del Carmen Sánchez y mi compañera Luz M^a Antequera.

A M^a del Carmen le agradezco su ayuda permanente e incondicional, sus sugerencias e ideas en la redacción de esta memoria y su inestimable conocimiento interno de la Universidad y de la gestión de sus datos.

A Luz, le agradezco su cordialidad, simpatía, paciencia y apoyo, gracias por escuchar mis paranoias durante todo este tiempo. Gracias por ser, más que una compañera, una amiga.

Gracias de corazón a mis compañeros, tanto de la administración como del Departamento de Estadística y Matemática Aplicada, por su cálida acogida, aliento y aprecio todos estos años; en especial, a Antonia que siempre me anima y a Jose Luis Alonso por su ayuda en lo referente a las aplicaciones de gestión de datos de la Universidad.

Agradezco el soporte económico por parte del Ministerio de Educación y Ciencia a través del proyecto TIN 2004-06204-C03-01, a los investigadores que desarrollaron (y continúan perfeccionando) el programa Elvira, herramienta fundamental en este trabajo, y al Servicio de Documentación del Consejo de Coordinación Universitaria, gracias a cuya eficiente gestión he podido recopilar abundante bibliografía sobre indicadores universitarios.

Gracias a mis amigos, Beltrán, Gil, Joaquín y Paco, por su paciencia durante mi encierro.

Por último, un agradecimiento especial a Mico, mi gata, por tanto cariño y buenos momentos en estos diez años juntas.

Índice general

INTRODUCCIÓN	1
Objetivos de la memoria	7
Organización por capítulos	8
1. INDICADORES DE RENDIMIENTO	9
1.1. Definición de indicador. Características y clasificación	12
1.1.1. Clasificación de indicadores	14
1.2. Críticas a los indicadores de rendimiento	17
1.3. Selección de indicadores	19
1.4. Catálogo de Indicadores del Sistema Universitario Español	28
1.5. Conclusiones	48
2. ELABORACIÓN DE LOS DATOS	49
2.1. Estructura de los datos.	49
2.1.1. Data warehouse.	53
2.2. Datos para el análisis del rendimiento de las asignaturas	56

2.3. Datos para el análisis del rendimiento en las titulaciones	60
3. INTRODUCCIÓN A LAS REDES BAYESIANAS	69
3.1. Conceptos básicos	69
3.2. Construcción de una red bayesiana a partir de una base de datos.	77
3.2.1. Aprendizaje usando técnicas de búsqueda. Algoritmo K2.	78
3.2.2. Aprendizaje de la estructura mediante detección de independen- cias. Algoritmo PC.	81
3.2.3. Estimación de las distribuciones condicionadas.	83
3.2.3.1. Método de Máxima Verosimilitud	83
3.2.3.2. Métodos basados en la Ley de Sucesión de Laplace	85
3.3. Propagación	86
3.3.1. Algoritmo HUGIN: un método de agrupamiento.	87
3.4. Extracción de perfiles	92
3.4.1. Abducción total	93
3.4.2. Abducción parcial	94
3.5. Programa Elvira	94
4. ANÁLISIS DE INDICADORES DE RENDIMIENTO. EL CASO DE LA UNIVERSIDAD DE ALMERÍA	97
4.1. Análisis por asignaturas.	98
4.1.1. Estudio de las asignaturas troncales y obligatorias.	114
4.1.2. Estudio de las asignaturas optativas.	121

4.2. Análisis por titulación.	128
4.3. Extracción de perfiles.	148
4.3.1. Perfil de asignaturas con mayor rendimiento.	148
4.3.2. Perfil de titulaciones con mayor rendimiento.	150
5. PREDICCIÓN EN ESTADÍSTICA UNIVERSITARIA	155
5.1. Redes bayesianas para regresión	156
5.2. El modelo MTE	158
5.3. El modelo de regresión Bayes ingenuo basado en MTEs	160
5.4. Selección de las variables a incluir en el modelo	164
5.4.1. Algoritmo para la construcción del modelo con selección de variables.	167
5.5. Estudio de casos	169
5.5.1. Discusión de los resultados	173
5.6. Conclusiones	175
6. CONCLUSIONES Y LÍNEAS ABIERTAS	183
6.1. Líneas de trabajo futuro	185
Bibliografía	187

INTRODUCCIÓN

La evolución de la enseñanza superior en España ha sufrido grandes cambios en las últimas décadas [39]. Podríamos diferenciar tres periodos en esta evolución:

El primer periodo comenzaría con la creación de las propias universidades hasta la década de 1970 aproximadamente. En este periodo, la educación superior se caracteriza por ser privilegio de unos pocos (en 1962 sólo un 4.3 % de los jóvenes correspondiente a la cohorte de edad de ingreso en la universidad, accedían a los estudios universitarios). Debido a esta circunstancia, el gasto público en esta enseñanza era mínimo.

El desarrollo económico y el progresivo aumento en el nivel de vida de la sociedad, así como en el número de jóvenes en edad universitaria, motivan, en un segundo periodo, una necesidad de invertir en educación superior: se crean nuevos centros universitarios, se amplían los existentes, aumentando la asignación de fondos públicos a la educación superior.

En la actualidad, nos encontramos en una tercera fase en la que la disminución en el tamaño de las cohortes de jóvenes (ver tabla 1) en edad universitaria provoca una estabilización (si no es descenso) del número de alumnos y esto a su vez, estanca, cuando no reduce, las aportaciones estatales o autonómicas al presupuesto universitario. En la tabla 2 puede apreciarse esta disminución del número de alumnos matriculados en las Universidades Españolas.

En este nuevo marco de estabilización financiera y humana, surge la necesidad de desarrollar mecanismos que mejoren la calidad de las universidades, así como de sistemas de información a través de los cuales las universidades puedan responder del

Edad	Años						
	1998	1999	2000	2001	2002	2003	2004
18	610.538	593.139	564.519	540.630	516.630	507.161	492.843
19	645.345	615.009	594.866	570.057	552.001	529.530	517.740
20	657.318	649.473	619.278	604.763	584.408	568.364	542.156
21	677.588	662.299	652.357	629.986	622.305	603.407	582.650
22	679.441	682.259	666.124	666.982	649.025	643.937	619.000
23	674.822	682.763	686.394	682.266	688.507	672.926	660.167
24	663.939	678.083	688.690	704.808	705.353	713.960	689.909
25	663.186	666.411	685.773	707.793	728.245	731.694	731.219
TOTAL	5.272.177	5.229.436	5.158.001	5.107.285	5.046.474	4.970.979	4.835.684

Tabla 1: Evolución de la población española en edad universitaria. Fuente INE

<i>1997-98</i>	<i>1998-99</i>	<i>1999-00</i>	<i>2000-01</i>	<i>2001-02</i>	<i>2002-03</i>	<i>2003-04</i>	<i>2004-05</i>
1.570.588	1.580.158	1.587.055	1.554.972	1.526.907	1.507.147	1.473.936	1.449.136

Tabla 2: Evolución del alumnado de las Universidades Españolas. Fuente INE

AÑO	GASTO (miles €)	AÑO	GASTO (miles €)
1992	3.084.794	1998	4.836.154
1993	3.339.657	1999	5.124.666
1994	3.480.046	2000	5.569.937
1995	3.844.928	2001	6.025.913
1996	4.235.273	2002	6.369.224
1997	4.758.481	2003	7.012.569

Tabla 3: Evolución del Gasto público total en educación superior. Fuente INE

uso que realizan de los recursos públicos asignados y de cuáles han sido los resultados de tal uso. Es aquí donde hacen su aparición los indicadores como soporte de los nuevos mecanismos de mejora de la calidad, toma de decisiones y de asignación de recursos (contratos programa). Estos modelos de financiación incentivos utilizan los indicadores como determinantes de una mayor o menor financiación pública para la universidad: un buen resultado, medido por los indicadores, implica un premio, en forma de una mayor financiación, mientras que un mal resultado implica un castigo reduciendo la subvención.

La introducción de estos sistemas de financiación ha suscitado numerosas críticas [37], la principal de ellas es la pérdida de la autonomía universitaria ante la intervención externa, a través de indicadores u otros mecanismos, en los objetivos que las universidades deben alcanzar.

Villareal [37] expone una triple argumentación en contra de la utilización de indicadores de rendimiento en la financiación pública:

- La validez o no del sistema de indicadores debido a la excesiva generalización que supone respecto a las circunstancias concretas de cada institución universitaria.

- La validez de los indicadores utilizados para reflejar los objetivos que se supone debe cumplir el sistema educativo superior.
- La capacidad técnica de la Administración para poner en práctica un sistema de indicadores que refleje efectivamente los objetivos a alcanzar.

Este autor, propone cuatro condiciones generales que deben cumplir los indicadores para que resulten operativos:

- Los objetivos que la Administración persigue deben ser claramente entendibles a partir de los indicadores seleccionados.
- Los indicadores deben ser adecuados y reflejar aquellos aspectos de los objetivos que se desea incentivar.
- Deben ser sencillos y calculables.
- Deben gozar de una cierta estabilidad temporal.

Las características que un sistema de indicadores debe cumplir para que sea fiable y válido han sido ampliamente estudiadas, como puede verse en [44, 39, 13].

Una vez establecido un sistema válido de indicadores, la siguiente cuestión a abordar es si los indicadores deben utilizarse con el mismo nivel de exigencia en todas las universidades. Dado que el potencial inicial entre las diversas universidades será distinto, aquellas universidades con una buena posición de partida recibirían una mayor financiación mientras que las universidades con problemas verán disminuir su financiación, perpetuando y aumentando las diferencias entre ellas. Para evitar esta situación, la exigencia de cumplimiento de los objetivos debe corresponderse con la posición inicial de cada institución.

Para terminar, habría que estudiar el volumen de financiación ligado al cumplimiento de objetivos, ya que la sensibilidad de los indicadores utilizados provocará variaciones en la financiación asignada a la universidad, por lo que a mayor volumen de financiación ligada a los indicadores, mayor será el efecto desestabilizador que puede provocar

en las instituciones. En la práctica este efecto puede evitarse limitando el volumen de financiación ligado a indicadores o limitando las variaciones anuales en la financiación de una institución.

En el ámbito andaluz, la financiación universitaria es establecida por el Pleno del Parlamento de Andalucía que, en su sesión de 7 de junio de 2001, aprobó el Dictamen de la Comisión de Educación relativo a la situación del Sistema Universitario de Andalucía. Como consecuencia de dicho Dictamen los máximos representantes de la Universidades Públicas Andaluzas y de la Consejería de Educación y Ciencia de la Junta de Andalucía suscribieron el Modelo para la Financiación de las Universidades Públicas de Andalucía para el periodo 2002-2006, estableciendo que parte de la financiación está vinculada al cumplimiento de determinados compromisos y a la consecución de resultados.

Con fecha 22 de julio de 2002 se firmó el Contrato Programa entre la Consejería de Educación y Ciencia de la Junta de Andalucía y la Universidad de Almería. El Contrato-Programa establece los objetivos de calidad y eficiencia, así como los indicadores para medirlos que, una vez evaluados, servirán de base para la aplicación de la financiación condicionada.

La Comisión de Seguimiento prevista en el Contrato Programa determina ocho indicadores básicos y comunes para todas las Universidades Públicas de Andalucía y cuatro indicadores específicos para cada Universidad. Los indicadores de cumplimiento básico son:

1. *Evaluación de la calidad de las titulaciones*: se define como el porcentaje de titulaciones evaluadas dentro del Plan Andaluz de la Calidad de las Universidades.
2. *Evaluación de la calidad en los programas de postgrado*: porcentaje de programas de postgrado evaluados respecto al total.
3. *Valoración de la actividad docente del profesorado*: porcentaje de profesores evaluados.
4. *Coefficiente de duración media de graduados respecto a la duración de las titulaciones*.

5. *Implicación del profesorado en la investigación científica*: porcentaje de sexenios concedidos sobre los sexenios totales posibles.
6. *Evaluación de la investigación científica según el P.A.I.*: Puntuación media ponderada de los grupos de investigación P.A.I. por macroáreas de conocimiento.
7. *Transferencia de resultados de investigación*: Ingresos medios anuales (por periodos de tres años), por funcionario doctor.
8. *Calidad y gestión de los servicios*: medida como el porcentaje de unidades de administración y servicios evaluados.

Y los indicadores específicos acordados con la Universidad de Almería son:

1. *Prácticas remuneradas de alumnos en empresas*: cociente entre el número de prácticas de empresa y el número de alumnos de la universidad.
2. *Inserción profesional en la Agencia de Colocación Universitaria*.
3. *Número de actuaciones con alumnos con discapacidad*.
4. *Número de ordenadores disponibles por estudiante*.

Así mismo, la Comisión de Seguimiento establece para cada ejercicio los valores base, mínimo y objetivo que les servirán a la Comisión para evaluar el grado de cumplimiento de los indicadores, en una escala de 1 a 5.

En función del grado de cumplimiento alcanzado, la Consejería de Innovación, Ciencia y Empresa asigna y tramita la financiación correspondiente. Si el grado de cumplimiento es nulo, no se obtendrá financiación afectada a los compromisos; si el grado de cumplimiento es igual o superior al nivel tres, le corresponderá la máxima financiación prevista. Entre dichos valores se aplicará la proporción correspondiente.

Si existiese detracción de la financiación inicialmente asignada, el 70 % de su montante total se distribuye entre las Universidades, siempre que obtengan un nivel de

cumplimiento superior al valor mínimo establecido; el 30 % de la financiación detraída se distribuye entre aquellas Universidades con un grado de cumplimiento superior a la media, teniendo en cuenta, en ambos casos, el grado de cumplimiento global alcanzado por la institución y en función del peso de la dotación del Contrato Programa.

Objetivos de la memoria

En este estudio nos centraremos en el cuarto indicador, el *Coefficiente de duración media de graduados respecto a la duración de las titulaciones*. El objetivo de reducir este indicador nos hace plantearnos qué variables influyen en el rendimiento académico de los alumnos. Para ello hemos realizado dos estudios: el rendimiento académico de los alumnos por asignaturas y el rendimiento académico por titulación.

En el primer estudio hemos recogido en 19 variables la información de 1.345 asignaturas impartidas en la Universidad de Almería durante el curso 2003-04. En el estudio de las titulaciones, hemos analizado los valores de 15 indicadores obtenidos en las titulaciones de la Universidad de Almería durante los cursos 2000-01, 2001-02, 2002-03 y 2003-04.

El problema que presenta un análisis estadístico tradicional para estas situaciones radica en el alto número de variables y de datos a manejar, así como en la heterogeneidad de las variables. Existen estudios previos sobre indicadores basados en correlaciones [21], pero tienen el inconveniente de que no proporcionan información sobre la estructura de dependencias entre los mismos ni un modelo operativo que permita realizar inferencias sobre el valor de un determinado indicador no observado dado el valor de otros indicadores. En este sentido, las redes bayesianas [35] son una herramienta muy útil, pues no sólo permiten obtener modelos de carácter predictivo, sino que también proporcionan información sobre las relaciones entre las variables del problema, lo que puede llevar a una mejor comprensión del mismo. La validez de las redes bayesianas para tareas de minería de datos ha sido analizada en [22], y su aplicación en temas de estadística universitaria ha quedado constatada en [32, 30, 31].

Organización por capítulos

La memoria comienza con un estudio de los indicadores de rendimiento: su definición, características y propiedades, objetivos de los indicadores y críticas que suscitan su uso. A continuación estudiaremos los indicadores elegidos por el Consejo de Coordinación Universitaria así como los problemas que estos indicadores presentan a la hora de su cálculo o para medir el objetivo para el que se proponen.

En el segundo capítulo comentaremos el estado en el que se encuentran los datos en la Universidad de Almería: organización de las bases de datos, el uso de integradores que conecten unas bases con otras así como las aplicaciones disponibles para la explotación de las bases de datos.

En el capítulo 3 haremos una introducción a las redes bayesianas y los principales métodos de propagación de probabilidades, terminando con el programa Elvira, que será la herramienta utilizada en los siguientes capítulos para modelizar el rendimiento académico del alumno.

El cuarto capítulo estará dedicado a los resultados empíricos de la investigación: el proceso de obtención y depuración de datos, la construcción de las distintas redes bayesianas y las conclusiones que de ellas se desprenden sobre el rendimiento académico del alumnado por asignaturas y titulación.

En el capítulo 5 construiremos un modelo de predicción basado en el empleo de redes bayesianas y distribuciones de tipo MTE. como caso práctico, aplicaremos dicho modelo para predecir el valor de indicadores universitarios, de forma que pueda ayudar al equipo de gobierno de la Universidad en la toma de decisiones sobre la política a seguir para la consecución de los objetivos de la institución.

Terminaremos la memoria recogiendo en el capítulo 6 las conclusiones sacadas del estudio.

Capítulo 1

INDICADORES DE RENDIMIENTO

Como comentábamos en la introducción, la Universidad Española ha llegado al final de un periodo de crecimiento y se enfrenta a una nueva situación en la que el número de estudiantes se estanca o decrece, por lo que las universidades han de competir por los estudiantes e instaurar mecanismos de gestión de recursos físicos y humanos así como procesos de evaluación y mejora de la eficiencia y eficacia de los programas educativos.

Es en este contexto cuando se hace patente la deficiencia en los métodos para la obtención de información fiable que permita la toma de decisiones. Por lo tanto, cualquier estrategia que quiera seguirse para promover la calidad de las universidades está supeditada a la construcción previa de un sistema de información fiable.

Mora [40], sintetiza las razones para desarrollar un sistema de indicadores en los siguientes puntos:

- Los procesos de mejora interna necesitan información estadística fiable y la construcción de indicadores que ayuden a comprender el funcionamiento de las propias instituciones.
- La introducción de procesos de Gestión de Calidad y de Gestión Estratégica

exigen disponer de información fiable y comparable con la de otras instituciones.

- El importante volumen de recursos dedicados a la educación superior exige que las universidades respondan sobre el uso que hacen de ellos mediante sistemas de información diáfanos.
- Los nuevos mecanismos de asignación de recursos, como los contratos-programa, exigen disponer de información fiable sobre la realidad de las instituciones.
- Los usuarios de las universidades (estudiantes y empleadores) deben disponer de información fidedigna que les permita tomar decisiones en base a criterios razonables.
- La movilidad de los titulados universitarios exige establecer sistemas de información sobre universidades. En este sentido, la Declaración de Bolonia de 1999, suscrita por los ministros de educación europeos puede ser el germen de un sistema europeo de información común.

El uso de indicadores de rendimiento suficientemente desarrollados, ayuda a entender la situación de los problemas con respecto a unas variables determinadas, a estudiar su evolución en el tiempo y poder hacer previsiones en un sistema tan complejo como es una institución universitaria.

En este sentido, De Miguel [11], en evaluación de programas, recomienda el uso de indicadores de rendimiento apoyándose en lo siguiente:

1. Aportan información válida, verificable y relativamente neutral.
2. Posibilitan establecer comparaciones en función de diversos criterios.
3. Permiten que los juicios de valor se apoyen sobre datos objetivos.
4. Facilitan el consenso entre la audiencia implicada.
5. Otorgan credibilidad a la toma de decisiones.

6. Añaden transparencia a todo proceso evaluado.

Así mismo, Cuenin [6] estudia, en el ámbito de asignación de recursos, el uso de indicadores como la *carga de trabajo*, el *número de horas impartidas*, *efectivos escolares*, *resultados de exámenes*, *tasas de matrícula*, *costos unitarios* o *naturaleza de las enseñanzas* para la asignación de presupuestos a las facultades y/o departamentos. Para la asignación de recursos para la investigación, hace referencia a indicadores de rendimiento como *subvenciones externas*, *plantilla de profesores*, *publicaciones de los docentes*, *tesis dirigidas*, *proyectos presentados*, etc. En cuanto a la asignación de recursos en personal, los indicadores más utilizados en la asignación de docentes son la *tasa de matrícula de alumnos* y la *carga de docencia* mientras que las asignaciones de recursos para personal no docente requiere ratios entre el personal docente y no docente.

También, en la planificación relativa a los estudiantes en la previsión de efectivos para los próximos años, se suelen emplear indicadores referentes a la admisión en el centro (*nivel de acceso*, *demandas de admisión* o *evolución de las tendencias del mercado*), el progreso de los alumnos (*tasa de éxito y de abandonos*), el número actual de estudiantes y su distribución en la universidad, el *número de egresados* y su origen geográfico.

En la gestión pedagógica se utilizan los indicadores como apoyo para la toma de decisiones. Un ejemplo de ello es la creación, modificación o supresión de planes de estudios, en donde suelen emplearse, además de los indicadores citados para la planificación de estudiantes, otros como la *inserción de los licenciados en el mercado de trabajo*, la *evaluación de los profesores* realizada por los estudiantes, la *carga de trabajo de los alumnos* o la *duración media de los estudios*.

Por tanto es clara la necesidad y la utilidad de los indicadores de rendimiento en la toma de decisiones de las instituciones universitarias. Pasemos ahora a definir qué es un indicador de rendimiento y qué propiedades tiene, para finalizar con un análisis de los indicadores propuestos por el Consejo de Coordinación Universitaria.

1.1. Definición de indicador. Características y clasificación

Debido al creciente interés por los indicadores de rendimiento, pueden encontrarse en la literatura numerosas definiciones. Como ejemplo de ello, algunas definiciones dadas son:

- Estadísticos que nos informan sobre la salud de un sistema [28].
- Un dato estadístico acerca del sistema educativo que revela algún aspecto sobre su rendimiento o salubridad [34].
- Datos empíricos, normalmente cuantitativos, sobre los recursos empleados y logros obtenidos en relación con los objetivos concretos de un programa [11].
- Procedimiento para obtener datos válidos y útiles a partir de los cuales se pueden tomar decisiones políticas [33].
- Ítem de información recogida de forma sistemática que permite establecer una información sobre un programa [11].
- Informes relativos a la manera en la que un sistema o parte de un sistema funciona [11].
- Estrategia o medio que permite presentar información relevante y oportuna sobre un programa a la audiencia interesada [1].

Concretando, Mora [39] llama **indicador de rendimiento** a toda medida objetiva, usualmente cuantitativa, de los logros de una institución o de todo un sistema de educación superior. Sin embargo Cuenin, [6], considera que el término *rendimiento* es interpretado con frecuencia por extensión, ya que muchos indicadores de rendimiento no miden, en realidad, rendimiento, por lo que aconseja emplear sistemáticamente el término *indicador* más que indicador de rendimiento.

La definición más concisa la da Dochy [13], que considera al indicador de rendimiento como una meta institucional, en vez de una medida de rendimiento. Dochy, en su definición, distingue entre *estadísticas de gestión*, *información para la gestión* e *indicadores de rendimiento*:

- *Estadísticas de gestión* son datos cuantitativos.
- *Información para la gestión* son datos cuantitativos o cualitativos que están relacionados entre sí y estructurados como información de dirección.
- *Indicadores de rendimiento* son datos empíricos de naturaleza cuantitativa o cualitativa que describen el funcionamiento de una institución. Estos datos se convierten en indicadores de rendimiento si expresan alguna meta perseguida.

Por ejemplo, el número total de alumnos matriculados en un curso académico es una estadística de gestión. El ratio graduados entre matriculados puede describirse como un dato de información para la gestión. Este ratio será un indicador de rendimiento si su mejora se persigue como una meta institucional.

Ante todas las definiciones dadas, se puede concluir que las características más relevantes que configuran un indicador de rendimiento [39, 43, 45] son:

1. *Validez*: el indicador mide lo que dice que mide.
2. *Relevancia*: El indicador debe ser útil para los interesados en los aspectos gerenciales de la actividad a la que se refiere el indicador.
3. *Fiabilidad*: las medidas son estables y replicables, de forma que individuos cualificados, usando los mismos datos y trabajando separadamente, deben llegar a los mismos resultados de los indicadores.
4. *Aceptabilidad institucional*: las personas afectadas por el uso del indicador deben haberlo aceptado, desde el punto de vista de su utilidad y validez.
5. *Economía* en la recogida de datos y su procesamiento, de forma que el beneficio de utilizar el indicador supere los costes de su cálculo.

1.1.1. Clasificación de indicadores

Al igual que ocurre con la definición de indicador, pueden encontrarse en la literatura diferentes clasificaciones para los indicadores. Expondremos a continuación algunas de las más relevantes:

Según el tipo de información que aportan [11, 13], los indicadores pueden clasificarse de la siguiente forma:

- Según las funciones de la educación superior se agrupan en:
 - *Indicadores de educación.*
 - *Indicadores de investigación.*
 - *Indicadores de servicios.*
- Dependiendo del tipo de medida,
 - *Indicadores cuantitativos.*
 - *Indicadores cualitativos.*
- Según el nivel de complejidad en la elaboración del indicador, diferenciamos entre:
 - *Indicadores simples:* que reflejan la magnitud de un fenómeno.
 - *Indicadores compuestos:* elaborados a partir de la interacción entre dos o más fenómenos (ratios, tasas o costos).
 - *Indicadores complejos:* contruidos para efectuar juicios de valor sobre el programa a evaluar, aunque pueden ser más difíciles de interpretar.
- En relación con los objetivos que pretendemos evaluar, se puede hablar de indicadores a partir de los cuales podemos:
 - Evaluar la *eficacia* de un programa.
 - Estimar la *eficiencia* en su aplicación.

<i>Criterio de Clasificación</i>	<i>Indicadores</i>
Función	educación investigación servicios
Tipo de medida	cuantitativos cualitativos
Forma de elaboración	simples compuestos complejos/índices
Objetivo a evaluar	eficacia eficiencia mejora
Parte del proceso evaluada	esfuerzo funcionamiento efecto
Fuente de la información	internos externos de investigación
Utilidad	Disponibilidad Publicidad Accesibilidad Utilización Adecuación Eficiencia Efectividad Impacto externo Aceptabilidad

Tabla 1.1: Clasificación de Indicadores

Indicadores de rendimiento

- Medir el grado de adecuación y *mejora* que el programa ha supuesto en relación con las necesidades planteadas.

Según la parte del proceso universitario que están evaluando, se pueden definir tres tipos de indicadores [24]:

- Indicadores de *esfuerzo* o de input, mide las entradas que la institución necesita para su funcionamiento.
- Indicadores de *funcionamiento* que miden la actividad de la institución.
- Indicadores de *efecto* o de resultados que miden los resultados obtenidos por el programa a evaluar.

Basándonos en la fuente de información de la que se extraen los datos [5]:

- *Indicadores internos*: se construyen a través de información obtenida en la propia institución.
- *Indicadores externos*: obtienen la información fuera de la institución.
- *Indicadores de investigación*: basados en datos de la actividad científica de la institución.

Utilizando como criterio la utilidad concreta que tienen los indicadores dentro del proceso evaluador [43], podemos clasificarlos en:

- *Indicadores de disponibilidad*: miden los servicios ofrecidos por la institución universitaria.
- *Indicadores de publicidad*: miden el conocimiento por parte de los usuarios de los servicios que ofrece la institución.
- *Indicadores de accesibilidad*: nivel de accesibilidad a los servicios ofrecidos.

- *Indicadores de utilización*: nivel de utilización y demanda potencial de los servicios existentes.
- *Indicadores de adecuación*: informan si es adecuada la relación cantidad-calidad de los servicios y las necesidades de las instituciones.
- *Indicadores de eficiencia*: comparan las entradas en la institución con el producto final obtenido.
- *Indicadores de efectividad*: comparan los logros obtenidos con los objetivos propuestos.
- *Indicadores de impacto externo*: miden el efecto y beneficio que las instituciones universitarias pueden ejercer sobre la sociedad.
- *Indicadores de aceptabilidad*: miden el grado de aceptación por parte de los usuarios de los servicios ofrecidos.

De forma esquematizada, podemos ver esta clasificación de los indicadores en la tabla 1.1.

1.2. Críticas a los indicadores de rendimiento

Si bien el uso de indicadores de rendimiento es recomendado por los organismos internacionales relativos al mundo de la educación, y utilizados en los planes de evaluación y contratos-programa, son muchas las críticas realizadas por parte de la comunidad universitaria. Tras más de veinte años de discusión, podemos resumir las principales críticas [16, 40, 39] en:

- Intentan medir cuantitativamente los resultados de la educación superior, algo que es esencialmente cualitativo. Para definir un conjunto de indicadores habría que llegar a un acuerdo sobre los objetivos culturales, económicos, tecnológicos y educativos de un sistema de educación superior y, posteriormente, trasladar

dichos objetivos a medidas cuantificables en la forma de indicadores de rendimiento, con pesos adecuados para cada objetivo, de modo que podamos medir la efectividad y la eficiencia de la institución.

- Intentan comparar algo que no es comparable, ya que los procesos de producción en la educación superior no aspiran a producir masivamente productos idénticos.
- Simplifican un proceso tan complejo y diverso como es el de la educación superior y la producción científica.
- Dadas las dificultades que rodean la definición, diseño y obtención de los indicadores, se tiende a calcular aquellos que son más fáciles de obtener o que están relacionados con partes del sistema universitario más fáciles de medir.
- Son ambiguos, a menos que sean usados en un contexto de diálogo adecuado.
- Cuando son utilizados por organismos externos preocupados esencialmente por las cuestiones financieras, puede que se primen las medidas de eficiencia y efectividad de las universidades a corto plazo, en detrimento de las metas que deban éstas tener a largo plazo.
- Existe el peligro de que los indicadores se conviertan en una meta por sí mismos.
- El uso de indicadores en la asignación de fondos puede provocar un círculo vicioso en el que malos resultados motivados por una falta de recursos dan lugar a una reducción de recursos.

A pesar de estas críticas, el uso de indicadores adecuadamente definidos e interpretados puede jugar un papel muy útil en la política, la gestión y la calidad de las universidades. Tras años de estudio y discusiones, hay una amplia bibliografía sobre el uso de indicadores, sus problemas, su validez y su interpretación. El problema siguiente que se nos plantea es el de encontrar un sistema de indicadores idóneo para las universidades.

1.3. Selección de indicadores

A estas alturas es imaginable que la construcción de un sistema de indicadores no es una tarea fácil; según De Miguel,[11], dicha construcción supone:

1. Establecer un marco conceptual o modelo teórico en el que aparezcan explícitos los fenómenos implicados así como las relaciones que se establecen entre ellos.
2. Identificar los componentes y/o dimensiones a evaluar.
3. Especificar, para cada dimensión o componente del modelo, los indicadores a través de los cuales puede ser observado.

Al ser un fenómeno estudiado por los expertos hace muchos años, existen numerosas propuestas de indicadores para el sistema de educación superior en la literatura:

En 1985, el *informe Jarratt* [25] recomendaba, para el sistema educativo inglés, el uso de un sistema de indicadores clasificados en tres categorías:

1. Indicadores de rendimiento interno:
 - Cuota de mercado que mantiene la institución respecto a solicitudes de matriculación de estudiantes en todo el país.
 - Tasa de graduación y tipos de grado conseguidos.
 - Atracción de estudiantes de doctorado y máster.
 - Tasa de éxito en los grados superiores y tiempo necesitado para completarlo.
 - Atracción de fondos para investigación.
2. Indicadores de rendimiento externo:
 - Aceptabilidad de los graduados en empleos.
 - Primer destino de los graduados.
 - Reputación de los graduados según juicio de evaluadores externos.

- Publicaciones y citas del profesorado.
- Patentes, inventos y consultorías.
- Premios y distinciones.
- Trabajos en congresos.

3. Indicadores de rendimiento operativo:

- Costes unitarios.
- Razón profesor/alumno; tamaño de las clases; disponibilidad de asignaturas a elegir.
- Carga docente.
- Disponibilidad de volúmenes en la biblioteca.
- Disponibilidad de ordenadores.

En 1987, Cuenin [5] realiza una investigación sobre los indicadores utilizados en un conjunto de universidades de la OCDE, clasificándolos por las cuestiones que pretenden valorar:

1. Indicadores para la asignación interna de recursos dentro de la institución:

a) Profesorado y departamentos:

- Carga Docente del profesorado.
- Número de estudiantes matriculados.
- Resultados de los exámenes.
- Razón profesor/alumno.
- Costes unitarios
- Tipo de asignatura.
- Fondos externos de investigación que se consiguen.
- Número de publicaciones.
- Número de empleados del departamento.

Indicadores de rendimiento

- Número de estudiantes de postgrado.
- Número de disertaciones doctorales.
- Número de proyectos de investigación.

b) Plazas de profesorado y de personal de administración y servicios:

- Razón profesor/alumno.
- Carga docente.
- Razón entre personal docente y no docente.

2. Indicadores de planificación institucional:

a) Estudiantes:

- Normas académicas de admisión.
- Número de solicitudes.
- Tendencias de las matriculaciones.
- Tasa de abandono.
- Número de graduados.
- Origen geográfico y étnico.

b) Fuentes financieras de la institución.

c) Planes de retiro del personal.

3. Indicadores de organización docente:

- Carga de trabajo que soportan los estudiantes.
- Valoración realizada por los alumnos de los cursos que reciben.
- Valoración de la enseñanza recibida por los alumnos ya egresados.
- Duración media de los estudios.
- Solicitudes de matrícula para cada curso y asignatura.

4. Indicadores de relación con los organismos externos:

a) Estudiantes: número de estudiantes clasificados por tipo de estudio y nivel.

- b) Profesores y personal de administración y servicios:
 - Profesorado clasificado por categorías.
 - Personal de administración y servicios clasificados por categorías.
- c) Aspectos académicos: número de graduados y ratio profesor alumno.
- d) Aspectos materiales: área construida y área no construida.
- e) Aspectos financieros: fondos recibidos clasificados por tipos.

En 1990, Dochy [12] propone una serie de indicadores para la educación superior clasificados en cuatro criterios:

- Funciones de la educación superior: docencia, investigación y servicios.
- Tipo de indicador: entrada, proceso o resultado.
- Medibilidad: cualitativa o cuantitativa.
- Validez: indicadores y variables, medidas del contexto y clarificación de los datos.

Como ejemplo, las tablas 1.2 y 1.3 muestran las propuestas de los autores para la función docente.

De Miguel [10], en 1991, propone 45 indicadores relativos a todos los factores y dimensiones que intervienen en los procesos de docencia:

- Infraestructuras:

Cuantitativos

- Tasa de espacio útil por alumno.
- Tasa de equipamiento especializado por alumno.
- Ratio de alumnos por profesor.
- Presupuesto por alumno.

<i>Indicadores de Personal</i>	
<i>VARIABLES CUANTITATIVAS</i>	<i>VARIABLES CUALITATIVAS</i>
Estructura de la edad Experiencia laboral Movilidad del personal	Calificaciones educativas Motivación docente
<i>Indicadores de Clima Educativo</i>	
<i>VARIABLES CUANTITATIVAS</i>	<i>VARIABLES CUALITATIVAS</i>
Tiempo medio y resultado de las innovaciones educativas	Filosofía educativa
<i>Indicadores de Infraestructura</i>	
<i>VARIABLES CUANTITATIVAS</i>	<i>VARIABLES CUALITATIVAS</i>
Número de ordenadores Nº de licencias de software	Calidad del espacio docente
<i>Indicadores de Egreso de estudiantes</i>	
<i>VARIABLES CUANTITATIVAS</i>	<i>VARIABLES CUALITATIVAS</i>
Tiempo medio entre graduación y el primer empleo % de desempleados tras x años % de doctores tras x años Media de publicaciones por Dpto. % de personal doctorado	Relevancia de la educación en relación con las actividades profesionales Oferta/demanda de plazas cualificadas
<i>Indicadores de Permanencia de estudiantes</i>	
<i>VARIABLES CUANTITATIVAS</i>	
% Alumnos que superan exámenes propedéuticos Obtención del doctorado Duración media del estudio por rama de enseñanza Tasa de abandonos	

Tabla 1.2: Indicadores de las variables educativas de entrada

<i>Indicadores</i>	<i>Variables Cualitativas</i>
<i>Personal</i>	Política de selección de personal.
<i>Oferta educativa</i>	Posibilidades de elección del currículum. Rango de elección de las principales materias. Oferta de áreas de estudio: posibilidades regulares, gratuitas y experimentales de formación a tiempo parcial, formación permanente y post-académica.
<i>Currículum</i>	Eficacia de los procedimientos para el desarrollo de la revisión de los programas. Claridad en los objetivos y contenidos del curso. Relevancia de los objetivos del curso en relación con la salida profesional. Acuerdo entre contenido y metas del currículum. Realización de las funciones del currículum. Corrección, plausibilidad, totalidad y nivel de los objetivos.
<i>Normativa</i>	Sistema institucional de control de la calidad. Acuerdos con otras instituciones referente a la distribución y concentración de tareas. Realización democrática de la normativa.

Tabla 1.3: Indicadores y variables educativas del proceso educativo

Cualitativos

- Buen mantenimiento de edificios e instalaciones.
- Equipamiento avanzado en laboratorios, bibliotecas y recursos.
- Programa de actividades formativas para el alumnado.
- Profesorado que dedica más tiempo a la docencia que a la investigación.

■ Planes de estudio:

Cuantitativos

- Porcentaje de estudiantes admitidos entre los preinscritos en relación con otras universidades.
- Porcentaje de alumnos procedentes de distintas regiones del país.
- Nota de acceso.
- Porcentaje de doctores que imparten docencia.

Cualitativos

- Enseñanzas acreditadas según expertos de otras universidades.
- Opiniones sobre la calidad de la enseñanza de antiguos alumnos, profesionales y empresarios.
- Prestigio del profesorado (reconocimientos).
- Vinculación de las enseñanzas con la sociedad y con la empresa.

■ Procesos de enseñanza:

Cuantitativos

- Ratio entre las horas de docencia teórica y práctica.
- Nº de horas dedicadas a la docencia en el curso en relación con las previstas.

Indicadores de rendimiento

- Tiempo dedicado por el profesor a la docencia (dentro y fuera del aula).
- Tiempo medio dedicado por el alumno al estudio (dentro y fuera del centro).
- Ratio de asistencia a las clases.
- Consultas de fondos bibliográficos y utilización de equipos.
- Tiempo dedicado a la coordinación y revisión de la Enseñanza.
- N° de actividades paralelas a la docencia organizadas en el centro.
- N° de alumnos que utilizan los servicios de apoyo a la docencia.

Cualitativos

- Estrategias de revisión de los planes docentes de los Centros y los Departamentos.
- Organización y funcionamiento de las comisiones de docencia.
- Revisión de los programas y materiales elaborados por el profesor.
- Métodos de enseñanza utilizados por el profesor en clase.
- Uso generalizado de medios en la docencia (laboratorios, ordenadores, medios auditivos, etc.).
- Revisión de los criterios y modelos de exámenes.
- Actividades organizadas por los alumnos dentro y fuera del campus.
- Programas y servicios de información y ayuda al estudiante.
- Sistemas establecidos para efectuar el seguimiento de la docencia y la revisión de las enseñanzas.

■ Resultados:

Cuantitativos

- Calificaciones obtenidas por el alumnado a lo largo de los cursos.

- N° de años empleados para la graduación en relación con los establecidos en el plan de estudios.
- Tasa de graduación.
- Rendimiento en pruebas de grado y exámenes profesionales (ej. MIR).
- Tiempo en que finalizan los estudios y desempeño de una actividad profesional.
- Tasa de costo por graduado.

Cualitativos

- Relación entre objetivos previstos y alcanzados.
- Prestigio de los titulados en la Institución.
- Programas de innovación en la docencia.
- Política en materia de perfeccionamiento del personal y el desarrollo de los recursos humanos.
- Mejoras insitucionales en función de los rendimientos.

En 1999, De Miguel, Caïs y Vaquera [9], realizan un ranking de las universidades públicas basándose en 83 indicadores.

En el año 2000, comenzó a aplicarse un sistema de indicadores diseñado por un grupo de trabajo creado por el antiguo Consejo de Universidades (actual Consejo de Coordinación Universitaria). Este sistema de indicadores es el utilizado en los planes de calidad para la evaluación de las titulaciones, los contratos- programa y otras actuaciones para la mejora de la calidad universitaria en Andalucía. Por ello, nos centraremos en el estudio de este sistema de indicadores y en las dificultades que plantea su cálculo.

1.4. Catálogo de Indicadores del Sistema Universitario Español

Este sistema de indicadores propuesto consta de 46 indicadores divididos en 7 bloques:

1. OFERTA UNIVERSITARIA

- ***Distribución interna de la oferta de titulaciones.***

Porcentaje de titulaciones oficiales ofrecidas por la universidad en cada una de las cinco ramas de enseñanza. Este indicador permite analizar la orientación disciplinar de la Universidad (técnica, humanística, multidisciplinar), no el grado de especialización.

- ***Distribución porcentual de la oferta de titulaciones.***

Porcentaje de titulaciones ofertadas por la universidad en cada una de las ramas de enseñanza sobre el número total de titulaciones incluidas en el Catálogo de titulaciones oficiales del Consejo de Universidades. Con esto, se pretende analizar en qué medida la diversificación actual de estudios del Catálogo se corresponde con la oferta efectiva existente y el grado de especialización o cobertura de la universidad en las distintas ramas.

- ***Adecuación de la oferta de estudios con relación al entorno.***

$$\frac{\frac{p_i}{\sum_i p_i}}{\frac{c_i}{\sum_i c_i}}$$

donde:

p_i número total de plazas en una rama i ,

c_i número total de contratos realizados en una rama i .

Con este indicador se analizaría cómo se ajusta la oferta de titulaciones a la demanda del mercado. El inconveniente de este indicador reside en el acceso

a las bases de datos del INEM para la obtención de los contratos realizados en cada rama. Además el CCU sugiere utilizar únicamente los contratos indefinidos, para evitar la problemática que supondrían los contratos temporales de renovación continuada.

Además, si el objetivo es estudiar la adecuación de la programación universitaria al mercado laboral, hay un desfase en el periodo de tiempo que se toma de referencia, ya que se comparan los contratos y las plazas ofertadas en un mismo año, pudiendo variar mucho la demanda del mercado laboral cuando los que cubren estas plazas ofertadas se graduen. Una alternativa para solucionar este problema podría ser comparar el ratio de la demanda laboral en una determinada rama, con el ratio de graduados en dicha rama.

2. DEMANDA UNIVERSITARIA

■ *Preinscritos en 1ª opción sobre la oferta de plazas.*

Mide el grado de adecuación de la oferta a la demanda de los estudiantes. Este indicador sólo tiene sentido cuando se calcula por titulación, pero no si se agrega por ramas debido a la gran heterogeneidad dentro de las ramas, como ocurre por ejemplo en CC. Sociales y Jurídicas.

Por otro lado, el cálculo del indicador se complica cuando la titulación no tiene límite de plazas. En esta situación, y para poder calcular el indicador por ramas, el CCU propone tomar la cantidad de 75 plazas en aquellas titulaciones sin límite de plazas ofertadas, sin embargo, en el caso en que la titulación tenga mucha demanda, esta solución supone un problema por sí mismo. Otra solución sería tomar 75 plazas en el caso en que el número de preinscritos no supere dicha cifra y el número de preinscritos en caso contrario.

■ *Nuevo ingreso en 1ª opción sobre el total de nuevo ingreso.*

Tiene como objetivo medir el grado de satisfacción de la demanda, por lo que sería más correcto el tomar como alumnos de nuevo ingreso aquellos que acceden por primera vez a través de preinscripción, es decir, excluyendo del denominador casos como traslados de expediente de otras universidades.

El problema que tiene este indicador es que tiende a disminuir en titulaciones sin límite de oferta de plazas, o con una oferta grande, ya que estas titulaciones tienden a recoger a los alumnos que no han sido admitidos en otras titulaciones.

- ***Nota media de acceso del 20 % superior.***

Con la nota media del percentil 80 de los alumnos se intenta medir la calidad del alumnado que ingresa en un estudio. El problema nos lo encontramos ante los distintos cupos de acceso a las titulaciones (*Pruebas de Acceso a la Universidad, Formación Profesional superior, Titulados, Mayores de 25 años...*), con la consiguiente variedad de escalas en las notas (los titulados poseen notas de acceso entre 1 y 4 mientras que los que acceden por pruebas de acceso o FP entre 5 y 10). Además, si lo que pretende medirse es el nivel a priori de los alumnos, ¿es correcto dar el mismo peso a la nota de acceso por PAU que por FP o Acceso de Mayores de 25 años?.

Una opción es calcular este indicador sólo para los ingresos en el cupo de *Pruebas de Acceso a la Universidad*, o bien calcularlo para cada una de las formas de acceso.

- ***Nota media de acceso.***

La nota media de acceso del alumnado de nuevo ingreso en el estudio se utiliza para analizar el nivel previo del alumnado que accede a la titulación. Se pueden aplicar los mismos inconvenientes que los detallados en el indicador anterior.

- ***Movilidad interautonómica de alumnos.*** Para estudiar la movilidad estudiantil en una comunidad autónoma y la capacidad de captación de estudiantes de otras comunidades, el indicador se define como

$$\frac{i_r}{i_0}$$

siendo:

i_r número de alumnos de nuevo ingreso procedentes de otras comunidades,

i_0 número de alumnos de nuevo ingreso matriculados en la comunidad.

Para realizar el estudio en una universidad, se definen i_r como el número de alumnos de nuevo ingreso en la universidad procedentes de otras comunidades e i_0 como el número de alumnos total de nuevo ingreso.

Al analizar este indicador debemos estudiar también la oferta de titulaciones de la comunidad/universidad, ya que una oferta pobre puede suponer que los estudiantes se desplacen para poder realizar los estudios que no pueden cursar en su comunidad/universidad.

■ ***Movilidad internacional de alumnos.***

Porcentaje de alumnos de la universidad que proceden de países extranjeros. Para su cálculo conviene determinar si se van a incluir los alumnos con becas de movilidad (Erasmus, Sócrates, etc...). Para determinar el alumnado que proviene de países extranjeros puede realizarse de dos formas:

- A través de la *Nacionalidad* del alumno: esta elección plantea el problema, cada vez más frecuente, de alumnos de nacionalidad extranjera y cuyas familias viven en España. La alternativa es, pues, tomar el
- *País del domicilio familiar* del alumno, que en nuestra opinión es la elección más acertada.

3. RECURSOS HUMANOS

En este apartado, se estudian indicadores del personal propio de la universidad (sin incluir centros adscritos).

Se entiende como *personal docente e investigador* (P.D.I.) a los funcionarios de los cuerpos docentes universitarios y el personal contratado.

- El profesorado funcionario comprende a aquellos profesores que acceden a su puesto a través de la correspondiente oposición, es decir, catedráticos, profesores titulares y profesores eméritos.
- El profesorado contratado abarca a ayudantes, profesores ayudantes doctores, profesores colaboradores, profesores contratados doctores, profesores asociados y profesores visitantes.

- No se incluyen los becarios.

El *personal de administración y servicios* (P.A.S) está formado por el personal funcionario o laboral no docente que configura la estructura administrativa, de gestión y de servicios de la universidad.

- ***P.D.I. a tiempo completo.***

Porcentaje de *PDI a tiempo completo* sobre el total de PDI.

Se define el personal docente e investigador *a tiempo completo* como el PDI que posee una dedicación docente de 8 o más horas semanales o su equivalente en créditos.

Este indicador mide la exclusividad del profesorado con la profesión académica.

- ***Porcentaje de P.D.I. doctor*** sobre el total de PDI de la universidad.

Nos ofrece una medida del potencial investigador de la Universidad.

- Porcentaje de ***P.D.I. funcionario*** en la plantilla.

Es un reflejo de la política estabilizadora del profesorado seguida por la universidad. Con la entrada de la LOU, debería modificarse este indicador incluyendo al profesorado contratado fijo.

- ***P.A.S./P.D.I.***

Relación entre el número total de P.A.S. y el número total de P.D.I.

Indicador del apoyo logístico que en recursos personales de administración y servicios cuenta la universidad para el mejor desarrollo de sus funciones docentes, investigadoras y de extensión cultural y servicios.

El Consejo de Coordinación Universitaria propone para medir este mismo objetivo el indicador:

- ***P.A.S./P.D.I. a tiempo completo.***

Consideramos que, dado el significado que el Consejo de Coordinación Universitaria atribuye a estos dos últimos indicadores, sería oportuno tener en cuenta un nuevo indicador que midiera la relación entre el P.A.S. y el número de alumnos.

Estos indicadores pueden calcularse por universidad y desagregados por ramas de enseñanza, titulaciones o departamentos. Sin embargo si quiere realizarse por centros habría que llegar a un consenso sobre la asignación del profesorado a cada centro:

- Bien un profesor se asigna a todos los centros donde imparta clase, o
- un profesor se asigna al centro donde tenga un mayor porcentaje de docencia. En este caso habría que considerar particularmente aquellos casos en los que un profesor imparta igual porcentaje en dos o más centros, en cuyo caso podría recurrirse al perfil del profesor.
- Análogamente, habría que consensuar la asignación del P.A.S. a cada una de las modalidades del nivel considerado, siendo especialmente complicado el tratamiento del P.A.S. de los servicios generales de la universidad.

En cualquier caso hay que tener especial cuidado si se hace una evaluación histórica de los valores de estos indicadores, pues hay que tener en cuenta el cambio en la docencia del profesorado de un curso a otro, que puede llevar a computar a un profesor en distintos centros cada año.

4. RECURSOS FINANCIEROS

Como en los indicadores del apartado anterior, los indicadores de este grupo excluyen los centros adscritos de la universidad.

Estos indicadores se salen del objetivo de este estudio, por lo que procederemos sólo a definirlos tal y como aparecen en el catálogo del Consejo de Coordinación Universitaria:

- ***Transferencias corrientes de las Administraciones Públicas sobre el total de ingresos corrientes.***

Relación porcentual de la participación financiera de las Administraciones Públicas en el conjunto de los ingresos corrientes de la Universidad. Es de mayor utilidad si se calcula en los ámbitos regional, nacional e internacional.

En el numerador se recogerían la financiación procedente de la Administración del Estado, de Organismos Autónomos, de la Seguridad Social, de Sociedades mercantiles públicas, de Comunidades Autónomas, de Corporaciones Locales y de diferentes Administraciones públicas no españolas. En el denominador, aparecerían los derechos reconocidos en concepto de tasas, precios públicos y otros ingresos; transferencias corrientes e ingresos patrimoniales que estén destinados a atender los gastos corrientes de la institución.

■ ***Precios públicos de enseñanzas de grado sobre el total de ingresos corrientes.***

Participación financiera de los derechos de matrícula de los alumnos de enseñanzas oficiales universitarias de primer y segundo ciclo en el conjunto de los ingresos corrientes de la universidad.

■ ***Ingresos generados por prestación de servicios sobre el total de ingresos corrientes.***

Participación financiera de los derechos generados por la prestación de servicios docentes distintos a las enseñanzas oficiales de primer y segundo ciclo, investigación aplicada y servicios de carácter cultural, residencial y otros, en el conjunto de los ingresos corrientes de la universidad.

■ ***Ingresos generados por la actividad investigadora sobre el total de ingresos no financieros.***

Participación financiera de los derechos generados por la investigación aplicada y por los proyectos de investigación, en el conjunto de los ingresos corrientes de la universidad.

■ ***Gastos de personal sobre el total de gastos corrientes.***

Indica el peso que los recursos humanos representan financieramente en la estructura de los gastos ordinarios de la Institución.

Antes de su cálculo habría que establecer si se contabilizan en el numerador los gastos generados por la prestación de servicios de empresas adjudicatarias, ya que el valor del indicador cambiará significativamente dependiendo

de la política de externalización de servicios de la universidad.

- ***Gastos de mantenimiento y conservación sobre el total de gastos corrientes.***

Indica el esfuerzo presupuestario de la universidad en el mantenimiento y renovación de sus instalaciones y equipamientos.

- ***Gasto corriente por alumno matriculado*** en enseñanzas de primer y segundo ciclo.

La utilidad de este indicador se manifiesta al realizar comparaciones institucionales.

- ***Gasto corriente por alumno matriculado corregido por la experimentalidad.***

Análogo al anterior, pero, en el denominador, se pondera a los alumnos de cada titulación por un coeficiente establecido por el Consejo de Coordinación Universitaria (1, 1.17, 1.30, 1.47, 1.51 y 1.56) dependiendo del grado de experimentalidad de cada titulación.

5. RECURSOS FÍSICOS

Este bloque de indicadores tienen como objetivo el estudio del grado de ocupación de los distintos espacios docentes.

Para el cálculo de los indicadores, el número de alumnos a considerar en el numerador dependerá del tipo de construcción de la universidad:

- *Modelo Departamental:* Campus donde los alumnos de las enseñanzas impartidas comparten los aularios, edificios de laboratorios, edificios departamentales, edificio de rectorado, edificio de gestión, edificio de restauración, etc. En estos casos, los datos harán referencia a cada uno de los campus de la universidad y el número de alumnos por puesto se calculara utilizando el número de alumnos de todas las titulaciones que comparten el campus.
- *Modelo Tradicional o de Facultad:* los datos hacen referencia a cada una de las facultades de la universidad y a las titulaciones que se imparten en ellas.

- ***Puestos en aulas.***

Mide el grado de ocupación de las aulas a través de la expresión:

$$\frac{N}{P_a}$$

siendo

N número de alumnos matriculados,

P_a número de puestos de uso simultáneo al día en aulas.

Para calcular el número de puestos simultáneo al día, se multiplica el número de puestos de uso simultáneo de cada aula por el número de turnos que hay en cada aula en un mismo día, estableciéndose un máximo de dos turnos de alumnos diferentes al día (mañana y tarde).

- ***Puestos en laboratorios.***

$$\frac{N_l}{P_l}$$

siendo

N_l número de alumnos matriculados en titulaciones con uso de laboratorios,

P_l número de puestos en laboratorios de uso simultáneo.

En este caso, se establece de forma general que el número máximo de turnos que pueden ocupar un laboratorio al día es 4 (dos de mañana y dos de tarde).

- ***Puestos en bibliotecas.***

$$\frac{N}{P_b}$$

siendo

N número de alumnos matriculados,

P_b número total de puestos en biblioteca.

Para el cálculo del número de puestos en biblioteca, sólo se tendrán en cuenta aquellas salas de libre acceso a libros con algún técnico del servicio de biblioteca a cargo, excluyéndose así las salas de estudio y de trabajo en grupo.

- ***Puestos de ordenadores.***

$$\frac{N}{P_c}$$

siendo

N número de alumnos matriculados,

P_c número total de puestos en sala de ordenadores.

Entendiéndose por *puesto* el PC o terminal exceptuando los del personal de la institución.

6. PROCESOS

Se agrupan en este apartado los indicadores que miden los distintos aspectos del proceso de enseñanza.

- ***Dedicación lectiva del alumnado en créditos.***

Se define como la media aritmética de los créditos matriculados por los alumnos de una titulación y da una aproximación de la carga de créditos del alumnado.

Para que el indicador sea representativo de la carga efectiva del alumnado, no han de considerarse los créditos matriculados y convalidados, adaptados, reconocidos etc.

- ***Prácticas requeridas.***

Cociente entre el número total de créditos prácticos requeridos (incluido prácticum) y el número total de créditos del plan de estudios.

Éste es un indicador de la importancia dada a las prácticas en un plan de estudios.

Esta definición sería correcta en el caso de planes de estudios que sólo contengan asignaturas troncales y obligatorias, ya que el plan de estudios, establece los créditos optativos y de libre configuración requeridos, pero el número de créditos prácticos dependerá de la elección de asignaturas que realice el alumno. En este aspecto no hay un método consensuado para el cálculo de los créditos optativos y de libre configuración. Las opciones para calcular los créditos prácticos optativos pueden ser:

- a) Realizar el cálculo sólo con los créditos troncales y obligatorios.
- b) Aproximar los créditos prácticos optativos, P_r , como el producto:

$$\begin{aligned}P_r &= \bar{P}_o \cdot \bar{O}_r \\ \bar{O}_r &= \frac{C_r}{\bar{C}_o}\end{aligned}$$

siendo:

\bar{P}_o número medio de créditos prácticos de las optativas ofertadas en el plan de estudios,

\bar{O}_r número medio de optativas requeridas,

C_r créditos optativos requeridos,

\bar{C}_o número medio de créditos de las optativas ofertadas

El cálculo de los créditos prácticos de libre configuración se complica al poder elegirse entre asignaturas de otros planes de estudios, por lo que, o bien no se tienen en cuenta en el cálculo del indicador, o se aproxima como en el caso de los optativos suponiendo que el alumno escoge toda la libre configuración entre la optatividad ofertada en su plan de estudios.

Sin embargo, al pretender obtener una medida de la importancia de las prácticas en el plan de estudios, consideramos más adecuado que este indicador se calcule únicamente para los créditos troncales y obligatorios.

- ***Optatividad requerida de la titulación.***

Cociente entre el total de créditos optativos y de libre configuración que debe cursar un alumno para la obtención del título y el total de créditos requeridos en el plan de estudios.

Es una medida de la flexibilidad curricular del plan de estudios.

- ***Oferta de optatividad de la titulación.***

Indica la oferta de optativas entre las que puede elegir el alumno para diseñar su currículo. Se calcula como el cociente entre el total de créditos optativos ofertados y el número de créditos optativos requeridos en el plan de estudios.

Hay que tener en cuenta que, al tomar como fuente el plan docente de un curso determinado, el valor del indicador podría variar de un año a otro (asignaturas bianuales, optativas que dejan de ofertarse, etc).

- ***Prácticas en empresa del plan de estudios.***

Relación entre el número de créditos que el alumno ha de realizar en empresas u organizaciones públicas (incluida la propia universidad) y el total de créditos prácticos requeridos en el plan de estudios.

Este indicador refleja la importancia que el plan de estudios concede a la realización de prácticas en empresa, pero plantea los problemas señalados en el indicador *Prácticas requeridas* en el cálculo del número de créditos prácticos a cursar por el alumno.

- ***Grupos grandes de teoría.***

Para medir el grado de masificación de los grupos de teoría de la titulación, se propone el porcentaje éstos con 80 o más alumnos sobre el total de grupos de teoría impartidos en la titulación.

Su utilidad decrece cuando vamos agregándolo por ramas de enseñanza, ciclos o universidad, debido a la gran diferencia entre el número de alumnos de las distintas titulaciones.

- ***Grupos pequeños de teoría.***

Informa sobre el grado de personalización en las asignaturas de la titulación, calculándose como el porcentaje de grupos de teoría con 20 alumnos o menos.

Como el indicador anterior, es útil al calcularlo por titulaciones pero va perdiendo utilidad conforme subimos de nivel de agregación.

■ ***Dedicación del Profesorado doctor funcionario al primer curso del primer ciclo.***

Definido como el cociente de dos ratios: la proporción de profesores doctores funcionarios en primer curso de la titulación entre la proporción de profesores doctores funcionarios en la titulación.

Con este ratio se pretende medir el nivel de especialización del profesorado que imparte docencia en primer curso en relación con el grado de especialización de los profesores de toda la titulación.

Este indicador así definido, puede inducir a interpretaciones erróneas, como la que aparece en el catálogo de indicadores del CCU, [8], es errónea:

Si una titulación obtiene un valor muy próximo a cero, nos estaría indicando que hay muy pocos profesores doctores funcionarios que imparten docencia en primero con respecto al conjunto de la titulación. Cuanto más se aproxime a uno, mayor será la semejanza entre el peso relativo de los profesores doctores funcionarios que imparten docencia en primero, respecto al conjunto de profesores doctores funcionarios que imparten docencia en la titulación. Cuando tome el valor de 1 estará indicando que todos los profesores doctores de la titulación dan clase a primero.

Claramente, el hecho de tomar el valor uno, no significa que todos los profesores doctores de la titulación dan clase a primero, sino que la proporción es la misma. Además en el ejemplo ilustrativo se afirma:

En la Universidad Pedro VI, el número de profesores doctores numerarios en primero es 5, el número de profesores en primero es 10, el número de profesores doctores numerarios en la titulación es 60 y el número de profesores en la titulación es 100. El estimador se calcula :

Proceso 6: $[(5/10) / (60/100)] = 0.83$, lo que indica que el 83 % de los doctores dan clase a primero.

La interpretación correcta de este ejemplo es que, en primer curso, el porcen-

taje de funcionarios doctores es un 17 % inferior al porcentaje de funcionarios doctores en la titulación.

Este indicador debería completarse con otro que midiera la especialización del profesorado en una titulación. Podría considerarse para ello el porcentaje de profesores doctores funcionarios que imparten clase en el estudio.

- ***Estudiante por profesor.***

Desagregado por titulaciones, ciclos o cursos, informa sobre el ratio profesor-alumno, dando una idea del grado de masificación de las clases.

Se calcula como el cociente entre los créditos matriculados por los alumnos y los créditos impartidos por el profesorado. Para que este cálculo sea correcto se deberían excluir los créditos convalidados, reconocidos, etc.

Su utilidad se ve limitada ante la imposibilidad de calcular los alumnos que realmente asisten a clase; una solución es utilizar una ponderación para los alumnos de primera matrícula, de segunda matrícula y de tercera o más matrículas, como se utiliza en las unidades de ordenación docente para el desdoble de grupos.

En su estudio sobre los indicadores utilizados en enseñanza, Cuenin [6], insinúa la posibilidad de ponderar a los alumnos según su nivel de estudio (primer ciclo o segundo ciclo) o según el coeficiente de experimentalidad de la titulación.

7. RESULTADOS

Bajo este epígrafe se reúnen los indicadores de medición de los resultados del proceso educativo en los estudiantes y profesores.

- ***Tasa de abandono (interrupción de estudios).***

Se define como la relación porcentual entre el número de alumnos de nuevo ingreso que debieron finalizar la titulación el curso anterior y que no se han matriculado en los dos últimos cursos (el evaluado x y el anterior).

Debida a esta definición, este indicador no puede calcularse para las titulaciones de Sólo Segundo Ciclo, ya que éstas sólo constan de dos cursos.

- ***Tasa de rendimiento.***

Cociente entre el número de créditos superados por los alumnos de un estudio y el número de créditos matriculados.

Para que el indicador sea representativo del grado de eficacia de la labor docente, se han de excluir de su cálculo los créditos adaptados, convalidados, reconocidos, etc.

- ***Tasa de éxito.***

Complementa al anterior en el sentido de analizar los resultados obtenidos en las pruebas de evaluación. Se define como el cociente entre el número de créditos superados por los alumnos de un estudio y el número de créditos presentados a examen.

Al igual que en el indicador anterior, no han de considerarse los créditos adaptados, convalidados, reconocidos, etc.

- ***Tasa de graduación.***

Es el porcentaje de alumnos de una cohorte de nuevo ingreso en primero que finalizan la titulación en los años establecidos en el Plan de Estudios (n).

Este indicador, presenta una serie de complicaciones que detallamos a continuación:

- No se tienen en cuenta en el numerador a los alumnos que terminan la titulación en $n - 1$ años (o menos), penalizando la tasa de graduación.
- No recoge la casuística de las posibles convalidaciones/reconocimiento de créditos de los alumnos de nuevo ingreso en primero que ya poseen una titulación universitaria.
- El caso de alumnos que compaginan los estudios con el trabajo disminuye la tasa de graduación de las titulaciones donde este tipo de alumno abunda (usualmente de Sólo Segundo Ciclo), a pesar de que el alumno apruebe mayoritariamente los créditos de los que se matricula.
- En titulaciones con proyecto fin de carrera, es muy usual que los alumnos realicen el proyecto una vez que han aprobado el resto de las asignaturas que forman el plan de estudios, por lo que, en el caso de haber

cursado un año por curso, el número de años que utilizan estos alumnos en graduarse es de $n + 1$. Esto hace que las tasas de graduación en las ingenierías estén próximas a cero. Para solucionar este problema, podrían tomarse los alumnos egresados en n años (es decir que superan los créditos necesarios para la obtención del título salvo el proyecto fin de carrera).

- Su agregación por ramas de enseñanza, centros o universidad hace que el indicador pierda utilidad debido a la gran heterogeneidad entre las titulaciones.

■ *Duración media de los estudios.*

Es la media aritmética del número de años que los alumnos matriculados en una titulación tardan en graduarse sin el proyecto fin de carrera.

Este indicador plantea los mismos problemas que el anterior, añadiendo además:

- Su cálculo se complica en el caso de alumnos que realizan un traslado de expediente desde otra universidad. ¿Deberían contabilizarse los años matriculados en el otro centro si lo que se intenta medir es el grado de eficacia productiva de una institución o estudio?, una respuesta afirmativa perjudica la tasa de la institución que recoge al alumno, mientras que si la respuesta es negativa, no se están teniendo en cuenta las posibles adaptaciones que pueda llevar a cabo el alumno en su expediente.
- La conocida sensibilidad de la media aritmética a los valores anormalmente extremos hace que alumnos que, anormalmente, superan en 5 o 6 veces el número de años establecidos en el Plan de Estudios (por ejemplo, por estar compatibilizando los estudios con un trabajo) afecten significativamente el valor del indicador. Para evitar este problema, podrían eliminarse los casos superiores al percentil 99.
- Carece de sentido calcular este indicador agregado por ramas de enseñanza, centro o universidad, al variar el número de años en el que se distribuye en plan de estudios (2 en las titulaciones de Sólo Segundo

Ciclo, 3 en las diplomaturas y 4, 5 o 6 en las licenciaturas, ingenierías superiores o arquitectura).

Como hemos comentado, uno de los indicadores de cumplimiento del contrato-programa de las Universidades Andaluzas, hace referencia a la duración de los estudios agregado por universidad. Para ésto, se propone una variación de este indicador, el *Coficiente de duración media de los estudios respecto al número de años de la titulación*. Se define, para los alumnos egresados en una titulación, como:

$$\frac{\sum n_i \cdot g_i}{n \cdot \sum_i g_i}$$

siendo

n_i número de años que tarda en graduarse un alumno,

n número de años establecidos en el plan de estudios,

g_i número de alumnos graduados en n_i años.

En el cálculo de este indicador pueden eliminarse los traslados de expediente y los expedientes adaptados, según se disponga o no del año en que comenzó los estudios en la universidad de procedencia, y los casos superiores al percentil 99, siempre y cuando se justifique debidamente el motivo de dichas exclusiones. Además, para evitar la influencia de los casos anómalos, a los alumnos que se gradúan en menor número de años que los previstos en el Plan de Estudios, se computan con un número de años igual a los previstos y a los egresados en más del triple de la duración prevista para el estudio, n , se computan con una duración de $3 \cdot n$ años.

Alemany en [2] justifica cómo el indicador *tasa de éxito* (y por extensión, la *tasa de rendimiento*, *tasa de graduación* y la *duración media de los estudios*) no puede ser utilizado de manera homogénea en todas las universidades/centros/titulaciones al prescindir de la tipología o perfil arquetípico de los alumnos. Estas conclusiones se basan en:

- a) Las *normativas de permanencia* son muy variadas entre las distintas universidades y la aplicación de normativas más o menos duras van a dar lugar a perfiles muy diferentes entre los alumnos de unas universidades y otras.
- b) La oferta de dos o más *turnos* de clase (mañana y tarde) facilita a los estudiantes la posibilidad de combinar sus estudios con el trabajo.
- c) La distribución por edades (porcentaje de alumnos con más de 25 años) o la distribución laboral de los estudiantes es significativamente distinta entre unas titulaciones y otras.

El autor propone la construcción de indicadores de rendimiento que se fundamenten en la idea de respetar el ritmo o itinerario elegido por el estudiante. Dicho ritmo se manifiesta en la matrícula del alumno, de forma que el número de créditos matriculados será mayor si se quiere finalizar el estudio en el periodo de escolaridad mínimo y menor si se quieren compatibilizar los estudios con otras actividades. En este caso se puede evaluar el rendimiento académico a través del porcentaje de créditos superados sobre el total de créditos matriculados por el alumno durante los años matriculados en la titulación.

■ ***Tasa de progreso normalizado.***

Aunque menos intuitivo, este indicador, equivalente a la *Tasa de graduación* y a la *Duración media de los estudios*, resuelve los problemas que planteaban los anteriores, midiendo de forma más precisa el progreso real de los estudiantes.

Se define como el cociente entre el número total de créditos superados por el alumno y el total de créditos matriculados a lo largo de sus estudios (incluyendo las veces que ha repetido).

Sin embargo, para Alemany [2], en el caso de que la normativa de permanencia penalice únicamente la extinción de convocatorias, el volumen de créditos matriculados por el alumno no pondrá de manifiesto el ritmo que éste se marca, por lo que propone evaluar el rendimiento académico en tasas

de graduación progresivas, es decir, el porcentaje de alumnos que en $n + 1$ (siendo n el número de años establecidos por el plan de estudios) años ha superado el 75 %, el 50 % y el 25 % del total de créditos de la titulación.

- ***Satisfacción con los estudios.***

Porcentaje de graduados que tres años después de terminar los estudios muestran un nivel razonable de satisfacción con la formación recibida.

Es sin duda una información valiosa para la institución, sin embargo, su forma de obtención (a través de una encuesta a los graduados), limita mucho su significación:

Por un lado, y debido a la Ley de Protección de Datos, sólo puede enviarse la encuesta a los alumnos que dieron su autorización para que la universidad se pusiera en contacto con ellos una vez terminados sus estudios.

Por otro, el sesgo de no respuesta podría llevar a resultados no significativos del indicador.

- ***Satisfacción con el empleo.***

Porcentaje de graduados que tres años después de finalizar los estudios están empleados en un trabajo razonablemente satisfactorio.

Presenta los mismos inconvenientes que el indicador anterior.

- ***Tasa de participación en proyectos de investigación.***

Porcentaje de profesores que participan en proyectos científicos competitivos (Programa Marco Europeo, Plan Nacional o Planes regionales con procesos rigurosos de evaluación)

- ***Proporción de sexenios.***

Para medir el nivel investigador del profesorado de la universidad se recurre al porcentaje de sexenios concedidos sobre el número de sexenios posibles de la universidad.

El cálculo de este indicador plantea un serio problema al intentar establecer el número de sexenios posibles. El CCU [8], entiende como posibles el número de años entre la lectura de la tesis y la fecha actual, con la condición de que sea funcionario, dividido por seis.

Esta es una aproximación muy amplia puesto que, aunque para poder obtener un sexenio de investigación el profesor debe pertenecer a los cuerpos docentes universitarios y ser doctor, en la evaluación se considera el periodo desde la entrada como profesional en la universidad, incluyendo el periodo como becario de investigación, por este motivo puede darse el caso de que el indicador supere el 100 %. Además puede plantearse el inconveniente de que las universidades no tengan informatizado el dato de la fecha de expedición del título de doctor.

Como comentábamos en la introducción, este indicador aparece en el contrato-programa como el número 5 (implicación del profesorado en la investigación científica). Debido a los problemas de homogeneidad entre las universidades en el cálculo del número de sexenios posibles, en el año 2005 se llega al consenso de establecer como universo el de los profesores doctores de los cuerpos docentes universitarios y definir el número de sexenios posibles como la parte entera de la mitad de los trienios reconocidos a partir de la funcionarización en dichos cuerpos entre dos. Esta fórmula puede entenderse de dos formas:

$$\begin{aligned} \text{sex. posibles} &= \sum_i E \left[\frac{t_i}{2} \right] \quad \text{o bien} \\ \text{sex. posibles} &= E \left[\sum_i \frac{t_i}{2} \right] \end{aligned}$$

siendo t_i los trienios reconocidos del profesor i .

La primera expresión da lugar a un menor número de sexenios posibles que la segunda, no habiendo ninguna indicación o consenso sobre cuál se debe coger, lo que repercute al realizar una comparación interinstitucional del valor del indicador.

Este acuerdo sigue siendo una aproximación porque no está teniendo en cuenta el periodo total que se puede someter a evaluación, lo que implica que el indicador pueda superar el 100 % en aquellas universidades en las que, por la política de estabilización del profesorado seguida, los profesores fun-

cionarios tienen una amplia experiencia investigadora previa a la obtención de plaza como funcionario de los cuerpos docentes universitarios.

- ***Producción de doctores.***

Se define como el porcentaje de tesis leídas en el último quinquenio sobre el número de doctores existentes en la universidad.

En el numerador se están contando todas las tesis leídas en una universidad, tanto si el director de la misma pertenece a la universidad, como si pertenece a otra institución. Por otro lado, el numerador no contempla aquellas tesis dirigidas por doctores de la universidad y leídas en otras universidades. Por esto, el indicador es correcto sólo si su objetivo es medir la producción de doctores realizada por la institución, pero no como medida de la producción de tesis que realizan los profesores doctores de la universidad.

1.5. Conclusiones

En este capítulo hemos querido aproximarnos al estudio de los indicadores universitarios, revisando los principales conceptos y argumentos, a favor y en contra de su uso, encontrados en la bibliografía.

Este camino nos ha conducido al catálogo de indicadores del CCU, utilizado actualmente como base de las evaluaciones de las titulaciones, contratos-programa y otras medidas de la calidad de las instituciones universitarias españolas.

Por último, hemos intentado manifestar los problemas que se plantean en el cálculo de estos indicadores, proponiendo posibles soluciones.

Capítulo 2

ELABORACIÓN DE LOS DATOS

Como hemos comentado en la introducción, la universidad se enfrenta a nuevos desafíos para lo cual necesita un sistema que le proporcione la información necesaria para la toma de decisiones de manera rápida, efectiva, ordenada y concreta.

En este capítulo introduciremos cómo se gestiona la información en las universidades, tomando como ejemplo la Universidad de Almería, para posteriormente definir los indicadores objetivos a análisis en los capítulos posteriores.

2.1. Estructura de los datos.

Los procesos diarios y las necesidades que surgen en la gestión de la información de la universidad son los que marcan las aplicaciones que gestionan los datos. En muchas universidades, como la Universidad de Almería las bases de datos principales se agrupan en tres grandes aplicaciones:

- HOMINIS: gestión y administración del personal de la Universidad.
- SOROLLA: asuntos económicos (contabilidad, presupuestos y tesorería).
- UXXI Académica: gestión del alumnado (matriculaciones, actas, expedición de

Gestión de Personal	⇒	HOMINIS
Asuntos Económicos	⇒	SOROLLA
Gestión de alumnos	⇒	UXXI Académica
Recursos docentes	⇒	UXXI Académica

Tabla 2.1: Grandes aplicaciones de gestión de datos en la UAL.

títulos, becas, etc.) y de los recursos docentes (ordenación docente, grupos, planes de estudio, horarios, etc.)

El objetivo es que las bases de datos de estas tres grandes aplicaciones estén integradas en una sola, de forma que todos los datos estén sincronizados. Pero esto no siempre ocurre, en la Universidad de Almería, actualmente sólo se encuentran integradas las aplicaciones UXXI Académica y HOMINIS, mientras que SOROLLA no se interrelaciona con las otras dos.

El proceso de integración se produce al nivel de datos: personas, áreas de conocimiento, departamentos, campos de profesorado (contratos, plazas, dedicación, cuerpo, categoría y escala, etc) de forma que cualquier modificación en un campo se transmite automáticamente a todos los módulos que hagan uso de dicha información. Las modificaciones en las bases de datos y aplicaciones no integradas, deben hacerse a mano, a través de comunicaciones oficiales, con el consiguiente desfase en la información.

Los datos utilizados en esta tesis han sido extraídos de la aplicación UXXI Académica. Esta aplicación se construye sobre 1.029 tablas y está compuesta de 11 módulos principales que interactúan entre sí, como muestran la figura 2.1.

- **Planes de estudio:** grafos de los distintos planes de estudio, mantenimiento de las asignaturas (carácter, duración y créditos por curso académico), oferta de la libre configuración para cada uno de los planes en cada curso académico,

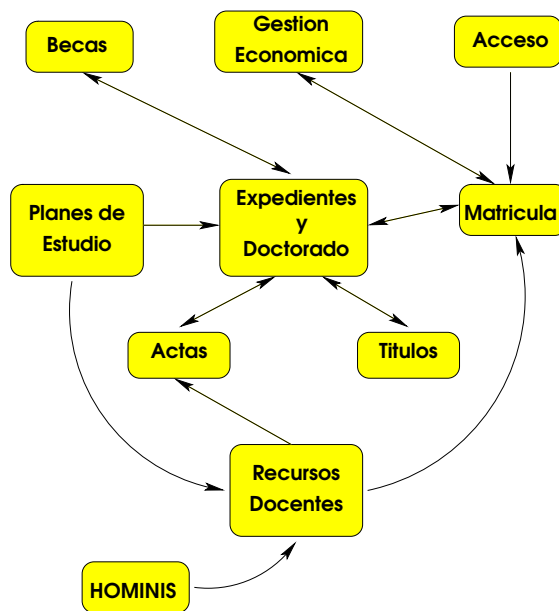


Figura 2.1: Relaciones entre los módulos de UXXI.

requisitos de los alumnos de un plan para la consecución de los distintos nodos (nodo de primer ciclo, de enlace de segundo ciclo, de finalización, etc).

- **Expediente:** expediente del alumno desde que éste inicia el estudio, acceso, convalidaciones, traslados, emisión de certificados y demás labores de gestión relacionadas con el historial académico del alumno.
- **Gestión económica:** gestión de los pagos de matrículas, certificados y títulos.
- **Actas:** tramitación y cierre de las actas de las asignaturas de un curso.
- **Matrícula:** realiza la matriculación del alumno en la secretaría y establece citas previas para los alumnos que realizan automatrícula. Controla la ocupación de los grupos de las asignaturas y la capacidad máxima para los alumnos que se matriculan como libre configuración; gestiona la matrícula del alumno (anulaciones, tanto de matrícula como de asignaturas, traslados a otros centros, etc).
- **Recursos docentes:** gestión de la ordenación docente, oferta de los grupos de



Figura 2.2: Aplicación UXXI Académica.

las actividades asociadas a cada asignatura.

- **Doctorado:** recoge las funcionalidades contenidas en los módulos de planes de estudio, actas, expedientes, matrícula y recursos docentes, adaptadas a la casuística de los programas de doctorado y de sus alumnos. Además gestiona las tesis doctorales.
- **Becas:** tratamiento y gestión de las solicitudes de becas de titulaciones oficiales de primer y segundo ciclo, tanto del Ministerio como las becas propias de la Universidad.
- **Titulos:** tramitación de las solicitudes de títulos.
- **Horarios:** gestión de horarios de todas las actividades ligadas a las enseñanzas regladas y no regladas, así como cualquier otra actividad organizada por la Universidad.

La aplicación tiene otros tres módulos, de **Acceso**, **Estudios propios** y **Extensión universitaria** que, actualmente, no se encuentran en explotación.

Fuera de estas tres aplicaciones principales, existen secciones y unidades, como Investigación, Relaciones internacionales, Docencia virtual, Enseñanzas propias, Formación del profesorado, Fomento al empleo, Infraestructuras, etc, que trabajan con bases de datos propias, usualmente en Access. Esto dificulta la explotación de los datos cuando deben establecerse correspondencias entre las distintas bases de datos. Por ello, la Universidad de Almería está implantando un data warehouse que relacione todas las bases de datos.

2.1.1. Data warehouse.

Un Data warehouse, a partir de ahora DWH, es una colección de datos en la cual se encuentra integrada la información de una Institución y que se usa como soporte para el proceso de toma de decisiones gerenciales.

El DWH de la Universidad de Almería se explota a través de Oracle Discoverer, integrado dentro de una plataforma Oracle 10g. Está formado por 11 *áreas de negocio*. Un área de negocio es un juego de información relacionada con un objetivo de negocio común. En nuestro caso, las áreas de negocio son:

- Gestión económica
- Infraestructuras
- Recursos humanos
- Selectividad
- Investigación
- Plan de ordenación docente
- Gestión académica



Figura 2.3: Áreas de negocio del DWH.

- Preinscripción
- Calidad
- Prácticas de empresa
- Plan de financiación

Dentro de cada área de negocio se definen *carpetas*, tal como muestra la figura 2.4. Las carpetas almacenan detalles sobre grupos de información relacionada y se corresponden con las tablas de la base de datos, vistas o combinaciones de tablas y vistas.

Una carpeta está compuesta por elementos (figura 2.5). Un elemento es la unidad básica que el usuario puede seleccionar, puede ser una visión lógica de una columna de la base de datos o bien, un cálculo a partir de varias columnas de la base de datos.

Los elementos de un área de negocio se relacionan a través de las *jerarquías*; su objetivo es el de permitir al usuario navegar entre elementos.



Figura 2.4: Carpetas del área de negocio Gestión Académica.

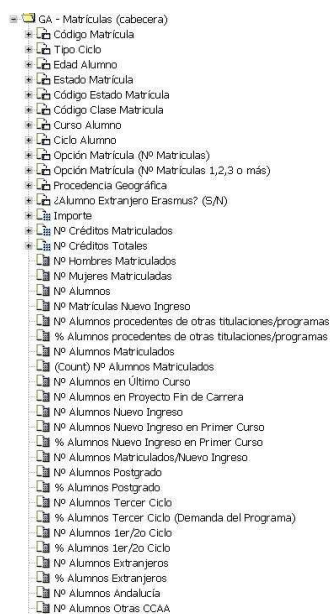


Figura 2.5: Elementos de la carpeta Matricula.

Una vez conocida cómo se estructura la información en nuestra Universidad, como se interrelaciona y las herramientas para explotarla, pasemos a exponer los criterios y definiciones seguidas para la construcción de las bases de datos objeto de nuestro análisis.

2.2. Datos para el análisis del rendimiento de las asignaturas

El objetivo es crear un perfil del rendimiento académico del alumnado en las asignaturas de la Universidad de Almería. Para ello creamos una base de datos con 1.345 asignaturas impartidas durante el curso 2003-04. Sólo han sido consideradas las asignaturas con docencia, esto es, no hemos recogido datos de asignaturas sólo con derecho a examen. También han sido excluidas asignaturas como los proyectos fin de carrera de las ingenierías, las prácticas de enseñanza de las distintas especialidades de magisterio, los practicum de Psicología, Psicopedagogía, Turismo y Diplomado en Gestión y Administración Pública, o los grupos de prácticas asistenciales en la Diplomatura de Enfermería, puesto que la naturaleza de estas asignaturas hacen que no puedan ser comparadas en el mismo marco que el resto.

Otro aspecto a tener en cuenta es que no hemos considerado a los alumnos que, habiéndose matriculado en una asignatura, la cursan en otra universidad a través de alguna beca de movilidad (Programas Erasmus, Seneca, etc.)

En las asignaturas de la titulación Licenciado en Derecho existe la casuística de asignaturas con alumnos con la calificación *Incompatible*. Se trata de asignaturas en las que los alumnos han de tener aprobada una o más asignaturas anteriores, requisito imprescindible para poder superarla. Se permite al alumno matricularse, pero la calificación será de *Incompatible* hasta que no haya superado estos requisitos. En estos casos hemos eliminado los alumnos en estas circunstancias, ya que influyen negativamente en la tasa de rendimiento.

En este marco hemos recogido información sobre 20 variables que pretenden describir al alumnado matriculado en la asignatura, características propias de la asignatura y del profesorado que la imparte. Dichas variables son:

- **Titulación:** se refiere a la titulación donde se financia la asignatura.
- **Carácter** de la asignatura: lo hemos resumido en dos valores: OB (asignaturas troncales y obligatorias) y OP (Optativas y Libre Configuración Específica). Los datos recogidos corresponden a 732 asignaturas troncales y obligatorias y 613 asignaturas optativas y de libre configuración específica.
- **Nº de alumnos por grupo de teoría.**
- **Nº de alumnos por grupo de práctica:** cuando los créditos prácticos de la asignatura se diversifican en distintas actividades (pizarra, ordenador, laboratorios) se han tomado los grupos de la actividad con mayor número de créditos. Si la docencia de la asignatura es compartida por dos áreas con distinto número de grupos, hemos tomado el mayor de ellos.
- **Nº de profesores distintos** que imparten la asignatura, ya sean créditos teóricos, prácticos en pizarra, laboratorios o en aulas de informática.
- **Evaluación del profesorado:** nota media obtenida por los profesores en los cuestionarios de evaluación, con puntuaciones de 1 a 5. Las asignaturas no evaluadas se han codificado con el valor 0, siendo así fácilmente identificables.
- Porcentaje de **créditos impartidos por profesores doctores.**
- Porcentaje de **créditos impartidos por profesores permanentes.**
- Nota **media de la nota de acceso** de los alumnos matriculados en la asignatura. Para unificar las notas de acceso de los distintos cupos, reescalamos las notas del cupo de titulados, con rango de 1 a 4, a un rango de 1 a 10 según la siguiente

expresión:

$$y = \begin{cases} 2 \cdot (x - 1) + 5 & \text{si } 1 \leq x < 2, \\ 1.5 \cdot (x - 2) + 7 & \text{si } 2 \leq x < 3, \\ 1.5 \cdot (x - 3) + 8.5 & \text{si } 3 \leq x \leq 4 \end{cases}$$

- **Percentil 80 de la nota de acceso** de los alumnos matriculados en la asignatura.
- Porcentaje de **alumnos repetidores** en la asignatura.
- Porcentaje de **alumnos a tiempo completo** en la asignatura. Se ha tomado como definición de alumno a tiempo completo la dada por el Consejo de Coordinación Universitaria:
Son aquellos alumnos que en los planes estructurados en créditos, se matriculan en 50 o más créditos y en los planes no estructurados en créditos al menos en cuatro asignaturas.
- **Tasa de éxito:** relación porcentual entre el número de alumnos que superan la asignatura y el número total de alumnos presentados a examen.
- **Tasa de rendimiento:** relación porcentual entre el número de alumnos que superan la asignatura y el número total de alumnos matriculados.
- **Tasa de aprobado:** relación porcentual entre el número de alumnos que obtienen la calificación "Aprobado" en la asignatura y el número total de alumnos presentados a examen.
- **Tasa de notable:** relación porcentual entre el número de alumnos que obtienen la calificación "Notable" en la asignatura y el número total de alumnos presentados a examen.
- **Tasa de Sobresaliente:** relación porcentual entre el número de alumnos que obtienen la calificación "Sobresaliente o Sobresaliente de Honor" en la asignatura y el número total de alumnos presentados a examen.

- **Media de convocatorias utilizadas** para superar la asignatura.
- **Índice de utilización de convocatorias:** definido, sólo para repetidores, como el número de convocatorias utilizadas dividido entre el número de convocatorias a las que tienen derecho. Toma valores entre 0 y 1. Para calcular el índice de la asignatura, se calcula la media de los repetidores. En caso de que la asignatura no tenga repetidores, se toma como valor el 1.

Para el cálculo del número de convocatorias a las que un repetidor tiene derecho, utilizamos el número de matrículas del alumno en la asignatura y la convocatoria en la que supera la asignatura (Febrero, Junio, Septiembre o Diciembre). Para las asignaturas cuya convocatoria ordinaria es en febrero o junio y la extraordinaria en diciembre o septiembre, podemos aplicar la fórmula:

$$2 \cdot (n-1) + \begin{cases} 1 & \text{si el alumno aprueba en convocatoria extraordinaria de diciembre,} \\ 2 & \text{en otro caso.} \end{cases} \quad (2.1)$$

siendo n el número de matrículas realizadas.

Esta fórmula obtiene el resultado correcto para aquellas asignaturas cuyos exámenes extraordinarios son en diciembre; sin embargo, en la Universidad de Almería, las facultades de Ciencias Experimentales y de Ciencias Económicas y Empresariales celebran los extraordinarios de las asignaturas del primer cuatrimestre en junio, las asignaturas anuales y del segundo cuatrimestre en febrero y la convocatoria extraordinaria fin de carrera en diciembre. En estos centros, la fórmula a aplicar para obtener el número de convocatorias posibles es:

$$2 \cdot (n-1) + \begin{cases} 1 & \text{si el alumno aprueba en convocatoria extraordinaria de diciembre} \\ & \text{o en febrero (sea ordinaria o extraordinaria),} \\ 2 & \text{en otro caso.} \end{cases} \quad (2.2)$$

- **Calificación relativa media de la asignatura:** se define la calificación relativa de un alumno i en una asignatura, NR_i como:

$$NR_i = \frac{N_i - NF}{NM - NF} \quad (2.3)$$

donde:

N_i : calificación obtenida por el estudiante i en la asignatura

NF : calificación mínima de la asignatura

NM : calificación máxima de la asignatura

Para obtener la calificación relativa media de una asignatura, calculamos la media de las calificaciones relativas de los alumnos en dicha asignatura.

Este valor se distribuye entre 0 y 1. En el caso en el que $NM = NF$, definimos la calificación relativa media de la asignatura como 1. Si no hay alumnos presentados o aprobados, la definimos como 0.

2.3. Datos para el análisis del rendimiento en las titulaciones

Se recogen los datos de los principales indicadores definidos por el Consejo de Coordinación Universitaria de las titulaciones con egresados desde el curso 2000-01 al 2003-04.

- **Optatividad requerida:** relación entre el número total de créditos optativos y de libre configuración que ha de cursar el alumnado a lo largo de la carrera y el número total de créditos a cursar por el alumnado para obtener el título correspondiente. Este indicador da una medida de la *flexibilidad curricular* que el plan de estudios permite al alumnado para que diseñe sus propios estudios.
- **Prácticas:** relación entre el número total de créditos prácticos requeridos (incluyendo los créditos del practicum) y el número total de créditos incluidos en el plan de estudios en su conjunto. Como comentábamos en la sección 1.4, no se han considerado los créditos optativos ni de libre configuración, puesto que éstos pueden variar según las asignaturas escogidas por el alumno. Este indicador muestra la importancia de la oferta de docencia práctica en la titulación.
- **Optatividad ofertada:** refleja el grado de oferta que presentan las titulaciones y se define como la relación entre el número total de créditos optativos diferentes ofertados en la titulación y el número total de créditos optativos que ha de cursar el alumnado.

El principal problema que se nos plantea es cómo contamos la optatividad ofertada, puesto que algunas optativas pueden estar ofertadas en unos cursos y en otros no. La solución elegida es tomar la máxima optatividad ofertada en el periodo que abarca desde el curso $x - n + 1$ hasta el curso evaluado, siendo

x : el curso evaluado,

n : el nº de años que consta el plan de estudios.

Esta decisión está basada en el hecho de que, la mayoría de los alumnos egresados en el curso x , han debido estar, al menos, matriculados en los n cursos anteriores, por lo que en alguno de ellos habrá podido disfrutar de la máxima optatividad ofertada en dicho periodo. Además, esta solución evita el problema que se daría en el caso de planes en extinción que cohabitan con el plan en proceso de implantación.

- **Grupos grandes de teoría:** relación porcentual entre el número de grupos de teoría con 80 o más alumnos y el conjunto de todos los grupos de las distintas asignaturas. Cuanto más se aproxime a 1, mayor será el grado de masificación de los grupos de la titulación.
- **Grupos pequeños de teoría:** relación porcentual entre el número de grupos de teoría con 20 alumnos o menos y el conjunto de todos los grupos de teoría. Aporta información sobre el grado de personalización de los grupos de la titulación.

Para que estos dos indicadores anteriores sean representativos de los grupos a los que ha asistido el alumno en sus, al menos, n últimos años de estudio, hemos tomado la oferta de grupos de teoría en asignaturas troncales y obligatorias siguiendo una evolución por los cursos que un alumno de nuevo ingreso en $x - n + 1$ (siendo x el año académico en el que se gradúa y n el número de años del plan) iría cursando si se matriculara a curso por año académico. En la tabla 2.3 aparecen, según el año académico de la promoción y el número de años en el que se distribuye el plan de estudios, n , la oferta de grupos a considerar para cada curso del estudio.

El cálculo de ambos indicadores para un plan en una determinada promoción de graduados en el año académico x se ha realizado como:

- *Grupos grandes de teoría:* $\frac{\sum_{i=1}^n G_i}{\sum_{i=1}^n T_i}$
- *Grupos pequeños de teoría:* $\frac{\sum_{i=1}^n g_i}{\sum_{i=1}^n T_i}$

donde G_i , T_i y g_i quedan establecidos en la tabla 2.2.

- **Porcentaje de créditos impartido por profesores doctores:** aporta información sobre el nivel de especialización del profesorado que imparte docencia en la titulación.

Siguiendo el mismo criterio que para el cálculo de los indicadores *Grupos grandes de teoría* y *Grupos pequeños de teoría*, hemos tomado el profesorado del curso c de la ordenación docente del año académico $x - n + c$, siendo x la promoción

Nº de alumnos en grupos de teoría	curso 1º oferta $x - n + 1$	curso 2º oferta $x - n + 2$...	curso n -ésimo oferta x
Más de 80	G_1	G_2	...	G_n
Menos 20	g_1	g_2	...	g_n
Total grupos impartidos	T_1	T_2	...	T_n

Tabla 2.2: Cálculo de los grupos grandes y pequeños de teoría según curso.

estudiada, n el número de años en el que se establece el plan de estudios y c el curso correspondiente, como puede observarse en la tabla 2.3. Para construir el indicador, se calcula el cociente entre la suma de los créditos impartidos por doctores en los n cursos y el total de los créditos impartidos en los n cursos. Las asignaturas optativas no tienen curso asignado y los créditos impartidos por doctores podrán variar según la ordenación que se escoja, por lo que para el cálculo de este indicador sólo hemos tenido en cuenta asignaturas troncales y obligatorias.

- **Estudiantes por profesor:** se calcula como el cociente entre el número de créditos matriculados por los alumnos y el total de créditos impartidos por el profesorado en dicha titulación. Con este indicador se pretende dar una idea del nivel de masificación a través de la proporción de estudiantes por profesor. En los créditos matriculados por los alumnos no se han tenido en cuenta las convalidaciones, adaptaciones, reconocimiento de créditos, etc.

Análogamente al cálculo de los anteriores indicadores, hemos tomado la situación de asignaturas troncales y obligatorias en el año académico $x - n + c$, siendo x la promoción estudiada, n el número de años en el que se establece el plan de estudios y c el curso correspondiente de la asignatura (ver tabla 2.3).

Este indicador plantea dificultades en su definición cuando se aborda su cálculo para una promoción de un determinado estudio. En este caso, el indicador podría calcularse de distintas formas:

1. Calcular los ratios para cada uno de los n cursos de la titulación, c , en el año académico $x - n + c$ y posteriormente, hallar la media de los n ratios. Este método plantea dos problemas: el primero de ellos viene dado por las materias optativas, que no pueden ubicarse en un año académico o curso determinado; si consideramos, como en los otros indicadores, sólo materias troncales y obligatorias, nos enfrentaríamos a un segundo problema derivado del reparto desigual de las asignaturas de carácter troncal u obligatorio en los cursos del plan de estudios. Los primeros cursos tienen un mayor número de estas asignaturas que va descendiendo conforme aumentamos de curso, en algunos casos, en el último curso de la titulación sólo se imparte una asignatura troncal u obligatoria. Este problema puede resolverse calculando la media de los ratios de cada curso, $R(c)$, ponderada por el número de asignaturas que interviene en el cálculo de cada ratio, a_c :

$$I(x) = \frac{\sum_{c=1}^n R(c) \cdot a_c}{\sum_{c=1}^n a_c} \quad (2.4)$$

2. Calcular el indicador correspondiente a la promoción x de un plan de n años como el cociente:

$$I(x) = \frac{\sum_{i=1}^n \text{créditos matriculados en el año } (x - n + i)}{\sum_{i=1}^n \text{créditos impartidos en el año } (x - n + i)} \quad (2.5)$$

En la fórmula (2.5), al no concretar el curso en el que se matricula, están interviniendo otras promociones ajenas a la estudiada.

3. Introduciendo el curso (considerando sólo asignaturas troncales y obligatorias) en la definición anterior, eliminaríamos el problema citado, pero aumentaría la complejidad del cálculo del indicador:

$$I(x) = \frac{\sum_{i=1}^n \text{créditos matriculados en el curso } i \text{ en el año } (x - n + i)}{\sum_{i=1}^n \text{créditos impartidos en el curso } i \text{ en el año } (x - n + i)} \quad (2.6)$$

Por su sencillez, en este trabajo hemos aplicado la fórmula (2.4) para el cálculo del indicador.

- **Nota media de acceso:** contribuye al análisis de la calidad del alumnado que accede a la titulación. Para ello se calcula la media de las notas de acceso de los alumnos de nuevo ingreso en 1º de la cohorte correspondiente al curso estudiado, es decir, para la promoción x , se estudiarían las notas de acceso de los alumnos de nuevo ingreso en el curso $x - n + 1$, siendo n el número de años que conforman el plan de estudios.
- **Dedicación lectiva del alumnado:** mide la carga de créditos media que tiene el alumnado calculando la media de créditos en los que los alumnos de una titulación se han matriculado. Para un año académico determinado, x , viene dado por el cociente:

$$\frac{CMAT(x)}{N(x)} \quad (2.7)$$

donde

$CMAT(x)$ total de créditos matriculados en el año x ,

$N(x)$ número total de alumnos en el año x .

- **Tasa de abandono:** porcentaje de alumnos de una cohorte de nuevo ingreso que debieron finalizar la titulación el curso anterior y que no se han matriculado ni en ese curso ni en el anterior al evaluado (es decir, no se han matriculado en los dos últimos cursos). Para el curso x es:

$$\frac{ABN(x, x - 1)}{NI(x - n + 1)} \quad (2.8)$$

donde

$ABN(x, x - 1)$ número de alumnos no matriculados

en los dos últimos cursos x y $x - 1$,

$NI(x - n + 1)$ número de alumnos de nuevo ingreso en el curso $x - n + 1$.

Para los estudios de Sólo Segundo Ciclo se ha calculado como el porcentaje de alumnos que, habiéndose matriculado en el curso $x - 1$, no se han matriculado en el curso x .

- **Tasa de rendimiento:** relación porcentual entre el número total de créditos superados (excluidos adaptados, Convalidados, Reconocidos, etc.) por los alumnos en un estudio y el número total de créditos en los que se han matriculado. Utilizaremos este indicador como medida del grado de eficacia del alumnado y de la institución docente con relación a su actividad académica.
- **Tasa de éxito:** relación porcentual entre el número total de créditos superados (excluidos adaptados, Convalidados, Reconocidos, etc.) por los alumnos en un estudio y el número total de créditos presentados a examen. Complementa el indicador tasa de rendimiento y permite analizar los resultados alcanzados en las pruebas de evaluación.
- **Tasa de graduación:** indica la eficacia productiva de las diferentes instituciones en lo concerniente al grado de satisfacción de la demanda de acreditación académica que anualmente manifiestan los usuarios de los servicios docentes universitarios. Se calcula, para un curso x , como el porcentaje de alumnos de nuevo ingreso en el curso $x - n + 1$ que finalizan la titulación en los n años establecidos por el plan de estudios.
- **Duración:** coeficiente de duración media de los estudios respecto al número de años de la titulación, calculado como:

$$\frac{\sum n_i \cdot g_i}{n \cdot \sum_i g_i} \quad (2.9)$$

siendo

n_i número de años que tarda en graduarse un alumno,

n número de años establecidos en el plan de estudios,

g_i número de alumnos graduados en n_i años.

- **Tasa de progreso normalizado:** proporción entre el número total de créditos que ha superado un graduado y el número total de créditos de los que se ha matriculado a lo largo de sus estudios (incluyendo las veces que ha repetido).

Promoción	Años Plan	Cursos				
		1º	2º	3º	4º	5º
2000-01	2	1999-00	2000-01	-	-	-
	3	1998-99	1999-00	2000-01	-	-
	4	1997-98	1998-99	1999-00	2000-01	-
	5	1996-97	1997-98	1998-99	1999-00	2000-01
2001-02	2	2000-01	2001-02	-	-	-
	3	1999-00	2000-01	2001-02	-	-
	4	1998-99	1999-00	2000-01	2001-02	-
	5	1997-98	1998-99	1999-00	2000-01	2001-02
2002-03	2	2001-02	2002-03	-	-	-
	3	2000-01	2001-02	2002-03	-	-
	4	1999-00	2000-01	2001-02	2002-03	-
	5	1998-99	1999-00	2000-01	2001-02	2002-03
2003-04	2	2002-03	2003-04	-	-	-
	3	2001-02	2002-03	2003-04	-	-
	4	2000-01	2001-02	2002-03	2003-04	-
	5	1999-00	2000-01	2001-02	2002-03	2003-04

Tabla 2.3: Oferta de grupos a tomar para cada promoción, titulación y curso.

Capítulo 3

INTRODUCCIÓN A LAS REDES BAYESIANAS

Durante las dos últimas décadas, el uso de redes para representar modelos probabilísticos ha representado un importante avance dentro de la Estadística, al hacer posible la implementación del razonamiento probabilístico en modelos donde interviene un gran número de variables. Esto se debe a que la propia estructura del grafo representa las relaciones de independencia entre las variables del modelo, de manera que la obtención de distribuciones a posteriori dado que se conoce el valor de algunas variables, puede hacerse sin tener que calcular la distribución conjunta. Esta tarea de cómputo de las distribuciones a posteriori se conoce como *propagación de probabilidad*.

En este capítulo introduciremos las redes bayesianas, su definición y los principales métodos de aprendizaje de redes a partir de una base de datos.

3.1. Conceptos básicos

Definición 1. *Una Red Causal es un grafo dirigido acíclico, es decir, un conjunto de variables y un conjunto de arcos dirigidos entre dichas variables que representan una relación de relevancia o de dependencia directa.*

Dada una red causal, podemos establecer las siguientes relaciones:

- A es *padre* de B si hay un arco $A \rightarrow B$, análogamente también diremos que B es *hijo* de A .
- A es *antecesor* de B si podemos encontrar un camino que, partiendo de A , alcance a B . En esta situación también diremos que B es un *descendiente* de A .

Definición 2. Llamaremos evidencia sobre una variable a una información sobre el estado de dicha variable. Si esta información nos indica el valor que toma la variable, la llamaremos evidencia fuerte o instanciación, en caso contrario la llamaremos evidencia débil.

La conexión entre las variables de una red causal puede ser en serie, divergente o convergente.

- **Conexión en serie:** una observación sobre X_1 influirá en la certeza sobre X_2 y, a través de ésta, en la de la variable X_3 y así sucesivamente hasta X_n . Y viceversa, una evidencia sobre X_n influirá, a través de X_{n-1} en mi certeza sobre los padres de cada variables hasta X_1 (ver figura 3.1).



Figura 3.1: Conexión en serie en una red causal.

Pero si conocemos el valor de X_i , con $1 < i < n$ bloqueamos el camino y la evidencia no podrá pasar de X_1 a X_n ni al contrario; diremos entonces que X_1 y X_n están *d-separadas* por X_i .

- **Conexión divergente:** veamos primero el siguiente ejemplo

Ejemplo 1. ([26]):

El inspector de policía Smith espera la visita de Holmes y del Dr Watson que ya se retrasan, mirando por la ventana se pregunta si las carreteras estarán heladas ya que si así fuera, y sabiendo que son conductores pésimos, lo más seguro es que hayan tenido un accidente.

Su secretaria entra y le comunica que Watson ha tenido un accidente con el coche. Smith contesta -“entonces las carreteras deben estar heladas y Holmes también tendrá un accidente, iré a comer yo sólo”. La secretaria le informa que las carreteras no están heladas, entonces Smith decide esperar a Holmes diez minutos más.

El razonamiento del inspector Smith puede modelizarse como se muestra en la figura 3.2, donde la variable C representa el estado de las carreteras y H y W a los sucesos “Accidente de Holmes” y “Accidente de Watson” respectivamente.

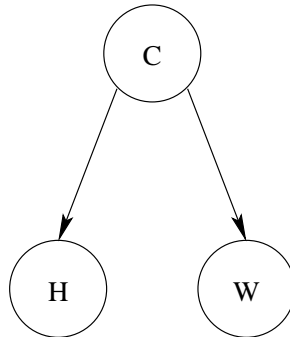


Figura 3.2: Ejemplo de conexión divergente en una red causal.

Cuando introducimos la información, “Watson ha tenido un accidente”, esto provoca un incremento en la creencia de Smith de que las carreteras están heladas, lo cual se propaga hasta H , haciendo que la creencia sobre que Holmes tenga otro accidente aumente. Pero cuando la secretaria le informa que las carreteras no están heladas, la información anterior ya no afecta a la creencia sobre H .

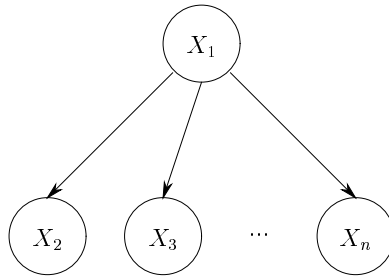


Figura 3.3: Caso general de conexión divergente en una red causal.

Generalizando esta situación, obtenemos la figura 3.3.

La influencia sobre mi creencia pasará, a través de X_1 , entre todos sus hijos, a no ser que conozcamos el estado de X_1 y diremos que X_2, X_3, \dots, X_n están *d-separadas* por X_1 .

- **Conexión convergente:** supongamos la siguiente situación:

Ejemplo 2. ([26])

Holmes está trabajando en su oficina cuando Watson le avisa que la alarma anti-robo (AAR) de su casa está sonando. Holmes coge su coche y se dirige rápidamente a su casa convencido de que algún ladrón (L) está en ella. Por el camino escucha por radio (R) la noticia de que ha habido un pequeño apagón (A) en la zona donde él vive, sabiendo que los apagones suelen provocar que salten las alarmas, Holmes da media vuelta y vuelve al trabajo.

El razonamiento de Holmes se esquematiza mediante la red de la figura 3.4.

Este es un ejemplo de conexión convergente (también se le conoce como **enlace cabeza a cabeza** y al nodo X_1 es un nodo cabeza a cabeza en el camino, que en general sigue el esquema de la figura 3.5.

Si no tenemos ninguna información sobre el estado de X_1 , los padres son independientes, es decir, la evidencia sólo puede transmitirse si el nodo cabeza a cabeza o uno de sus hijos ha recibido alguna evidencia.

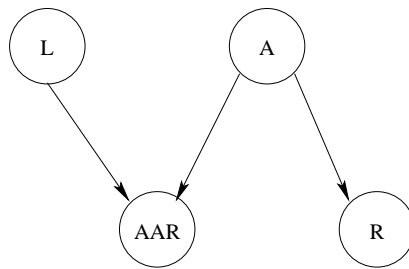


Figura 3.4: Ejemplo de conexión convergente en una red causal.

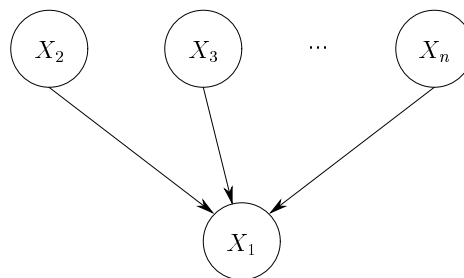


Figura 3.5: Caso general de conexión convergente en una red causal.

La importancia del concepto de independencia radica en que nos permite determinar qué información es relevante y cuál es superflua, pudiendo así esquematizar el conocimiento de forma que sólo sea necesario consultar las variables relevantes.

El criterio gráfico de independencia en redes causales es la **d-separación**.

Definición 3. *Dos variables X e Y en una red causal están d-separadas si en todos los caminos entre X e Y hay una variable intermedia V de forma que*

- *La conexión es en serie o divergente y V es conocida.*
- *La conexión es convergente y ni V ni ninguno de sus hijos han recibido evidencia.*

Si X e Y no están d-separadas se dicen *d-conectadas*.

Como se observa en los ejemplos anteriores, si X e Y están d-separadas, un cambio en la certeza de una no afecta en la certeza de la otra. Sin embargo, a pesar de que X e Y estén d-conectadas los cambios en la creencia de X no tienen por qué afectar en la de Y .

Denotaremos por $\langle X | Z | Y \rangle$ en un grafo dirigido acíclico G al hecho de que X e Y estén d-separados por Z , es decir, si todos los caminos entre X e Y están bloqueados por Z .

El bloqueo en la transmisión de la evidencia descrito está contemplado en la Teoría de la Probabilidad en el concepto de **Independencia Condicional**:

Definición 4. Sea (Ω, \mathcal{A}, P) un espacio probabilístico. Dadas tres variables X , Y y Z con $P(Y, Z) > 0$, se dice que X e Y son condicionalmente independientes dado Z si

$$P(X|Y, Z) = P(X|Z)$$

La condición $P(Y, Z) > 0$ no es relevante para nuestros propósitos ya que si, para ciertos estados de las variables, $Y = y$ y $Z = z$ tales que $P(Y = y, Z = z) = 0$, si Z toma el valor z , entonces $P(Y = y) = 0$ y la evidencia no aparecería.

Definiremos la relación de Independencia Condicional entre variables, $I(\bullet, \bullet, \bullet)$, como

$$I(X, Z, Y) \Leftrightarrow P(X|Y, Z) = P(X, Z)$$

En una red causal, podemos asignar números a los arcos para reflejar la *fuerza* de la relación entre las variables que une dicho arco.

En el marco de la Teoría de la Probabilidad, parece lógico pensar que, si X es padre de Y , la fuerza del arco que une X e Y sea $P(Y|X)$. Si Z también es padre de Y , entonces necesitaremos $P(Y|X, Z)$.

Además, para poder considerar un grafo dirigido acíclico, con un conjunto de distribuciones de probabilidad condicionadas asociadas a cada nodo, como una representación de una distribución de probabilidad conjunta, es necesario que las relaciones de independencia expresadas por el grafo sean válidas en la distribución de probabilidad .

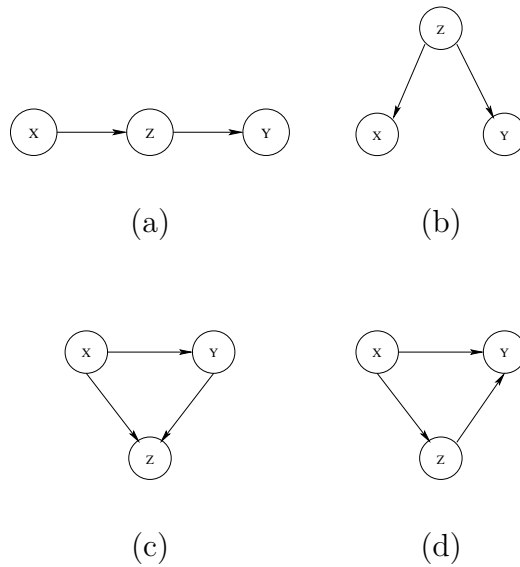


Figura 3.6: Distintos I-mapas del modelo $I(X, Z, Y)$.

Una posible representación gráfica para una distribución de probabilidad es un *I-mapa*:

Definición 5. Un grafo dirigido acíclico G es un I-mapa o mapa de independencias de una distribución P si toda relación de *d-separación* en G corresponde a una relación de independencia válida en el modelo P , es decir,

$$\langle X|Z|Y \rangle_G \Rightarrow I(X, Z, Y)_P$$

Un I-mapa garantiza que los nodos que están *d-separados* corresponden a variables condicionalmente independientes, pero puede ocurrir que independencias en el modelo sean dependencias en el grafo.

Definición 6. Un I-mapa se dice *minimal* si, al eliminar un arco o arista del grafo, éste deja de ser I-mapa.

Ejemplo 3. ([27])

Sea P una distribución de probabilidad con $I(X, Z, Y)$. Los grafos de la figura 3.6.(a) y 3.6.(b) son I-mapas *minimales*.

El grafo de la figura 3.6.(c) es un I-mapa trivial, pues no representa ninguna relación de independencia, y es minimal, ya que si eliminamos cualquier arco aparecen relaciones de independencia no contempladas por P .

El de la figura 3.6.(d) es un I-mapa no minimal ya que podemos eliminar el arco $X \rightarrow Y$.

Definición 7. Dada una distribución de probabilidad P , sobre un conjunto de variables X , una red bayesiana es un grafo dirigido y acíclico $G = (\Upsilon, \varepsilon)$, en el que

- A cada variable de X le corresponde un único nodo de Υ y viceversa.
- G es un I-mapa minimal de P
- Cada variable X_i tiene asociada una distribución de probabilidad condicionada a las variables $X_{\pi(i)}$, $P(X_i|X_{\pi(i)})$, donde $\pi(i)$ denota los subíndices de los padres de X_i en el grafo.

Proposición 1. (Regla de la Cadena)

Dada una red Bayesiana sobre un conjunto de variables, $X = \{X_1, X_2, \dots, X_n\}$, la distribución de probabilidad conjunta en X es el producto de las distribuciones de probabilidad condicionadas especificadas en la red, es decir,

$$P(X) = \prod_{i=1}^n P(X_i|X_{\pi(i)})$$

La regla de la cadena nos dice cómo encontrar una distribución de probabilidad a partir de una red bayesiana, pero, ¿es posible la situación recíproca?, es decir, dada una distribución de probabilidad, ¿podemos encontrar una red bayesiana que represente a P ? La respuesta a esta pregunta es afirmativa y dedicaremos la siguiente sección a la construcción de redes bayesianas partiendo de una distribución o de una base de datos.

3.2. Construcción de una red bayesiana a partir de una base de datos.

El problema al que nos enfrentamos es poder relacionar los datos **en bruto**, demasiados numerosos para comprenderlos fácilmente, con un modelo que pueda predecir o estimar valores de interés en situaciones o casos aún no observados. El método tradicional para obtener información de los datos es su análisis manual y su posterior interpretación, pero cuando el volumen de los datos aumenta de forma importante este tipo de análisis resulta demasiado costoso, por lo que se hace necesaria la automatización de este proceso.

En este apartado se ofrece una introducción a la construcción y aprendizaje automático de modelos gráficos considerando distintas metodologías y algoritmos que intenten resolver este problema.

Entenderemos por *aprendizaje* al conjunto de métodos semi-automáticos basados en la experiencia que intervienen en la construcción y/o posterior modificación de un modelo (llamado proceso de adaptación).

Los modelos gráficos constan de dos componentes: una cualitativa y otra cuantitativa. La componente cualitativa es un grafo que representa relaciones de dependencia e independencia (en el caso de redes bayesianas se trata de un grafo dirigido acíclico); la componente cuantitativa es una colección de parámetros numéricos que expresan la fuerza de las dependencias y miden nuestra incertidumbre (en redes bayesianas dichos parámetros serán distribuciones condicionadas de un nodo dado sus padres en el grafo).

Dadas estas componentes, podemos distinguir entre dos tipos de aprendizaje: el cualitativo (concerniente a la estructura de la red) y cuantitativo (especificación de las probabilidades condicionadas).

Ambos aprendizajes no pueden realizarse independientemente el uno del otro, es decir, para poder estimar las distribuciones de probabilidad condicionadas necesitamos conocer su estructura gráfica (necesitamos saber cuáles serán los padres de X_i para

calcular $P(X_i|X_{\pi(X_i)})$; además, para saber si en el grafo es necesario un arco entre dos nodos determinados, tendremos que calcular la intensidad de la relación entre tales nodos.

Comenzaremos estudiando métodos de aprendizaje de la estructura de la red.

Tenemos numerosos algoritmos para resolver este problema, todos ellos intentan encontrar una solución óptima en el conjunto de grafos dirigidos acíclicos que contienen las variables del problema. Estos algoritmos pueden clasificarse básicamente en dos grupos:

- Métodos basados en técnicas de búsqueda.
- Métodos basados en detección de independencias.

3.2.1. Aprendizaje usando técnicas de búsqueda. Algoritmo K2.

Estos métodos usan técnicas de búsqueda para obtener redes candidatas que se evaluarán siguiendo dos criterios:

- Por un lado, la red debe representar lo mejor posible al modelo original (los datos).
- Por otro, está la cuestión del tamaño o complejidad de la red.

Para evaluar cada red se utilizan métricas muy variadas, aunque las podemos englobar en tres principios:

- *Principio de Entropía*: buscan una red cuya entropía cruzada con los datos sea mínima. La entropía mide el grado de dependencia entre variables y estos métodos buscan redes que contengan conexiones entre las variables con alto grado de

dependencia. Un ejemplo de medida de entropía será

$$Dist(P, \tilde{P}) = \sum P(w) \log \frac{P(w)}{\tilde{P}(w)}$$

siendo P la distribución original y \tilde{P} la distribución sobre la red.

- *Descripción de Longitud Mínima:* buscan aquellas redes que minimicen la suma de las longitudes de codificación del modelo y de los datos dado el modelo, sin dejar de lado la exactitud del mismo. Una distancia para medir el tamaño de una red M es, por ejemplo,

$$Size(M) = \sum_{X_i \in U} Sp(X_i)$$

donde U es el conjunto de variables de la red y para cada variable X_i , $Sp(X_i)$ es el número de valores en $P(X_i|X_{\pi(i)})$.

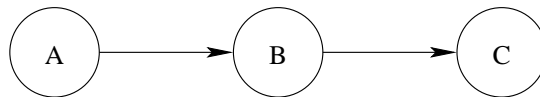


Figura 3.7: La red M del ejemplo 4.

Ejemplo 4. Dada la red M mostrada en la figura 3.7 con A , B y C variables binarias, el tamaño de M será:

$$Sp(A) = 2, Sp(B) = 4 = Sp(C) \Rightarrow Size(M) = 10$$

- *Métodos Bayesianos:* estos algoritmos buscan la estructura que maximiza la probabilidad de obtener una red condicionada a la base de datos disponible empleando:

$$P(Red|Datos) = \frac{P(Datos|Red) \cdot P(Red)}{P(Datos)}$$

Como los datos son siempre los mismos para cualquier red, el denominador es constante, $P(Red)$ es la distribución a priori de cada red candidata (que en muchos casos supondremos uniforme), por lo que $P(Red|Datos)$ depende únicamente de $P(Datos|Red)$, llamada *evidencia*, que es la verosimilitud muestral promedio.

A continuación describiremos el funcionamiento de uno de los algoritmos más conocidos en el aprendizaje de la estructura. Este algoritmo, llamado K2, utiliza una métrica bayesiana para calcular la distribución de probabilidad de una estructura G y una base de datos BD . Supuesto que los casos de la base de datos son independientes, que no hay datos en blanco y que las distribuciones de probabilidad de los parámetros de la red son uniformes, dicha métrica será

$$P(G, BD) = P(G) \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk!} \quad (3.1)$$

siendo

$$\begin{aligned} r_i &= \text{número de casos de la variable } X_i \\ x_{ik} &= \text{k-ésimo valor de la variable } X_i \\ q_i &= \text{número de casos de } X_{\pi(i)} \\ w_{ij} &= \text{j-ésimo valor de } X_{\pi(i)} \\ N_{ijk} &= \text{n}^\circ \text{ de casos en BD en los que } X_i = x_{ik} \text{ y } X_{\pi(i)} = w_{ij} \\ N_{ij} &= \sum_{k=1}^{r_i} N_{ijk} \end{aligned}$$

El término $P(G)$ puede obviarse, puesto que para simplificar la búsqueda exigimos la uniformidad de la distribución a priori sobre las distintas estructuras. También se ordenarán las variables, y el algoritmo recorre las variables según este orden y para cada variable se van incluyendo los padres de ésta que incrementen la probabilidad de la estructura resultante, es decir, que maximizan la función.

$$g(X_i, X_{\pi(i)}) = \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk!} \quad (3.2)$$

maximizándose así la probabilidad (3.1).

El pseudocódigo de este algoritmo es:

Dado un orden de las variables, sea V_i el conjunto de nodos anteriores a X_i y u el número máximo de padres de X_i .

Para $i = 1$, hasta n :

1. $X_{\pi(i)} = 0$
2. $P_v = g(X_i, X_{\pi(i)})$
3. $OK = \text{True}$
4. Mientras $OK = \text{True}$ y $|X_{\pi(i)}| < u$ hacer:
 - a) $Z =$ nodo de V_i que maximiza $g(X_i, X_{\pi(i)} \cup \{Z\})$
 - b) $P_N = g(X_i, X_{\pi(i)} \cup \{Z\})$
 - c) Si $P_N > P_V$ entonces
 - 1) $P_V = P_N$
 - 2) $X_{\pi(i)} = X_{\pi(i)} \cup \{Z\}$Sino $OK = \text{False}$.
5. Devolver $X_{\pi(i)}$

3.2.2. Aprendizaje de la estructura mediante detección de independencias. Algoritmo PC.

El objetivo de los métodos anteriores era encontrar una red que representara lo mejor posible a nuestra base de datos. Los métodos basados en test de independencia condicional lo que buscan son redes que representen la mayor parte de las relaciones de independencia entre las variables.

Por ello, tendremos que suministrar al algoritmo las relaciones de independencia condicional o bien una fuente de información para que pueda obtenerlas, como la distribución de probabilidad o una base de datos.

Si la información de entrada es una base de datos, el algoritmo obtendrá las relaciones de independencia aplicando test estadísticos de independencia condicional, lo

que influirá en la complejidad (el orden de un test depende del número de variables del conjunto al que se condiciona) y robustez (pues la detección fiable de independencias de orden elevado requiere gran número de datos).

Veamos la forma de actuar de estos métodos a través del algoritmo PC, uno de los más conocidos entre los que utilizan detección de independencias.

El algoritmo PC parte de la hipótesis de que todas las relaciones de independencia condicional del modelo se corresponden con relaciones de d-separación en el grafo y viceversa.

Los resultados fundamentales son:

Proposición 2. *Dado un grafo dirigido acíclico G ,*

- $X - Y \notin G \Leftrightarrow \exists S$ adyacente a X o a Y tal que X e Y están d-separados por S .
- Si $X - Y, Y - Z \in G$, y $X - Z \notin G$ entonces Y está en cualquier subconjunto de variables que d-separa a X y Z o no está en ninguno, en cuyo caso tendríamos $X \longrightarrow Y \longleftarrow Z \in G$

PC parte del grafo completo no dirigido y va eliminando aristas, primero las que unen nodos con independencia condicional de orden cero (independientes), luego las de orden uno (dada una variable), etc.

El pseudocódigo de este algoritmo es:

1. Formar el grafo completo no dirigido G
2. $n = 0$
3. **Mientras** $\forall X, Y$ adyacentes, con $|Ady(X, Y)| \geq n$ **hacer**:
 - **Repetir**
 - Seleccionar X, Y adyacentes tal que $|Ady(X, Y)| \geq n$
 - Seleccionar $S(X, Y) \subseteq Ady(X, Y)$ con $|S(X, Y)| = n$

- Si $S(X, Y)$ d-separa a X e Y , eliminar $X - Y$ de G y guardar $S(X, Y)$ hasta que todos los pares X e Y hayan sido comprobados.
 - $n = n + 1$
4. En el grafo G resultante, para cada terna X, Y, Z con $X - Y - Z \in G$ pero $X - Z \notin G$, orientar $X \longrightarrow Y \longleftarrow Z \Leftrightarrow Y \notin S(X, Z)$

3.2.3. Estimación de las distribuciones condicionadas.

Una vez que tenemos la estructura de la red, el grafo determina el conjunto de padres para cada variable. El problema ahora consiste en estimar los valores de las distribuciones de probabilidad condicionadas $P(X_i | X_{\pi(i)})$ a partir de la base de datos.

3.2.3.1. Método de Máxima Verosimilitud

Es el más simple y muy habitual. Consiste en estimar las distribuciones de probabilidad mediante las frecuencias relativas de ocurrencia de sucesos:

$$P(x_i | x_{\pi(i)}) = \frac{n(x_i, x_{\pi(i)})}{n(x_{\pi(i)})}$$

donde

- $n(x_{\pi(i)})$ es el número de casos en la base de datos en que las variables $X_{\pi(i)}$ toman e valor $x_{\pi(i)}$
- $n(x_i, x_{\pi(i)})$ es el número de casos en la base de datos en los que $X_i = x_i$ y $X_{\pi(i)} = x_{\pi(i)}$

Cuando el tamaño muestral es bastante grande el estimador de máxima verosimilitud tiende al valor verdadero, pero da problemas cuando la base es pequeña; veamos estos problemas mediante un ejemplo.

Ejemplo 5. ([7]) Supongamos que tenemos cuatro variables binarias y la base de datos:

X_1	X_2	X_3	X_4
1	0	0	1
0	1	0	1
0	1	1	0
0	1	0	0
1	0	0	1
1	1	0	1

Para la que la red óptima es la de la figura 3.8.

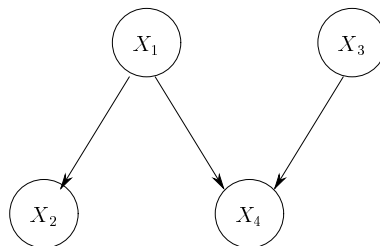


Figura 3.8: Red bayesiana relativa a la base de datos del ejemplo 5.

Las probabilidades serán

	$X_1 = 0$	$X_1 = 1$
$X_2 = 0$	0	$\frac{2}{3}$
$X_2 = 1$	1	$\frac{1}{3}$

	$X_3 = 0$	$X_3 = 1$
$X_1 = 0$	$(\frac{1}{2}, \frac{1}{2})$	$(1, 0)$
$X_1 = 1$	$(0, 1)$?

Observando las tablas vemos el primer problema: el estimador no está definido si el número de datos de cierto tipo es cero, esto ocurre al calcular $P(X_4|X_1 = 1, X_3 = 1)$.

El otro problema se debe a que el estimador se sobreajusta a los datos, es decir, como en nuestra base de datos no hemos encontrado configuraciones en las que $X_1 = 0$ y $X_2 = 0$, el estimador asigna una probabilidad cero a $P(X_2 = 0|X_1 = 0)$ aunque lo más seguro es que sea más alta. Otro caso análogo lo tenemos en $P(X_4 = 0|X_1 = 1, X_3 = 0)$ y $P(X_4 = 1|X_1 = 0, X_3 = 1)$.

Para evitar estos problemas se plantean los siguientes métodos de enfoque bayesiano:

3.2.3.2. Métodos basados en la Ley de Sucesión de Laplace

Si en una muestra de N casos encontramos K casos que verifican una determinada propiedad Q , entonces la probabilidad de que el siguiente caso que observemos exhiba la misma propiedad es $\frac{K+1}{N+|Q|}$, donde $|Q|$ es el número de alternativas posibles para la propiedad Q

Éste es el enunciado de la ley de sucesión de Laplace, que, aplicada a estimar $P(x_i|x_{\pi(i)})$ daría lugar al método:

$$P(x_i|x_{\pi(i)}) = \frac{n(x_i, x_{\pi(i)}) + 1}{n(x_{\pi(i)}) + |X_i|}$$

siendo $|X_i|$ el número de estados de la variable X_i .

Propiedades:

1. Si no hay datos, coincide con la distribución uniforme
2. Si la muestra es suficientemente grande, este estimador coincide con el de máxima verosimilitud

Ejemplo 6. *Las distribuciones de probabilidad del ejemplo anterior serían, aplicando este método:*

$P(X_2 X_1)$			$P(X_4 X_1, X_3)$		
	$X_1 = 0$	$X_1 = 1$		$X_3 = 0$	$X_3 = 1$
$X_2 = 0$	$\frac{1}{5}$	$\frac{3}{5}$	$X_1 = 0$	$(\frac{1}{2}, \frac{1}{2})$	$(\frac{2}{3}, \frac{1}{3})$
$X_2 = 1$	$\frac{4}{5}$	$\frac{2}{5}$	$X_1 = 1$	$(\frac{1}{5}, \frac{4}{5})$	$(\frac{1}{2}, \frac{1}{2})$

Como puede observarse ya no tenemos el problema de $n(x_{\pi(i)}) = 0$ y se ha disminuido los efectos del sobreajuste a los datos.

3.3. Propagación

Una vez construida la red bayesiana, se quiere ver cómo cierta información sobre algunas variables de la red impactan en las distribuciones de probabilidad del resto de variables.

Se llama **propagación de probabilidad** en una red bayesiana al cálculo de la distribución a posteriori de ciertas variables de interés dado que se conoce el valor de algunas otras variables.

Una forma de calcular estas distribuciones sería marginalizar la distribución conjunta; sin embargo manejar directamente la distribución conjunta puede presentar problemas de eficiencia y de almacenamiento (para el caso más sencillo de una red bayesiana

donde las n variables sean binarias tendríamos que almacenar 2^n valores de probabilidad).

Esto motivó el desarrollo de algoritmos que, aprovechando la independencia condicional entre las variables (puesta de manifiesto en la estructura de la red mediante la d-separación), realizan la propagación mediante cálculos locales.

A pesar de que proporcionan resultados exactos, el problema en sí es NP-duro, pues el tiempo de respuesta de estos algoritmos crece exponencialmente con el número de variables de la red. Sin embargo pueden encontrarse estructuras sencillas, como grafos encadenados (un nodo tiene como máximo un solo padre y un solo hijo), árboles (cada nodo sólo tiene un padre) y poliárboles (sólo hay un camino entre dos nodos cualesquiera), en los que el tiempo de respuesta es polinomial.

Puede consultarse más información sobre estos métodos en [23] y [26].

3.3.1. Algoritmo HUGIN: un método de agrupamiento.

Definición 8. *Dado un conjunto de variables X_I , llamaremos potencial a cualquier función $f : \Omega_{X_I} \rightarrow \mathbb{R}_0^+$.*

Un ejemplo de potencial es la distribución de probabilidad.

Definición 9. *Dado un potencial f , notaremos por $s(f)$ al conjunto de índices de las variables para las que está definido f .*

En una red bayesiana de variables $X = \{X_1, X_2, \dots, X_n\}$, denotaremos la distribución condicionada de X_i dados sus padres en la red $X_{\pi(i)}$ a

$$f_i(x) = f_i(x^{\downarrow i}, x^{\downarrow \pi(X_i)}), \quad \forall i \in \{1, \dots, n\}, \quad \forall x \in \Omega_{X_{s(f_i)}}$$

y la distribución conjunta vendría expresada como

$$p(x) = \prod_{i=1}^n f_i(x^{\downarrow s(f_i)}), \forall x \in \Omega_{X_N},$$

donde $x^{\downarrow s(f_i)}$ denota las coordenadas del vector x cuyos índices son $s(f_i)$.

Definición 10. Dado un potencial f definido sobre un conjunto de variables X_I , y dado $J \subseteq I$, se define la marginalización de f sobre X_J (o la eliminación de las variables con índices en $I - J$) como un nuevo potencial $f^{\downarrow J}$ definida sobre J

$$f^{\downarrow J}(x) = \sum_{\substack{y \in \Omega_{X_I} \\ y^{\downarrow J} = x}} f(y), \forall x \in \Omega_{X_J}$$

Definición 11. Dado r potenciales f_1, \dots, f_r , definidos sobre los conjuntos de variables con índices I_1, \dots, I_r respectivamente, se define su combinación como un potencial definido sobre el conjunto de variables con índice en $I = \bigcup_{i=1}^r I_i$ y dado por

$$f(x) = \prod_{i=1}^r f_i(x^{\downarrow I_i}), \forall x \in \Omega_{X_I}$$

Definición 12. Sea un potencial f definido sobre el conjunto de variables con índice en I , sea $J \subset I$ y $x_0 \in \Omega_J$. La restricción de f a x_0 es un potencial definido sobre las variables con índices en $I - J$

$$f^{R(X_J=x_0)}(x) = f(y), \forall x \in \Omega_{X_{I-J}} \text{ tal que } y \in \Omega_{X_I} \text{ con } y^{\downarrow I-J} = x \text{ e } y^{\downarrow J} = x_0$$

El algoritmo Hugin se basa en la transformación de la red bayesiana en una estructura llamada **árbol de grupos**, es decir, árboles en los que los nodos son conjuntos de variables de la red bayesiana original y los arcos se etiquetan con *separadores*, que consisten en la intersección de los nodos adyacentes. La propagación se realiza entonces mediante un paso de información a través de los arcos del árbol de grupos. El árbol de grupos se ha de construir a partir de una red bayesiana verificando:

1. Cada familia de la red (se entiende por *familia* al conjunto que forman una variable y sus padres en la red) se encuentra en, al menos, un nodo del árbol.
2. Propiedad de la unión: para cada par de nodos del árbol C y C' con

$$S = C \cap C' \neq \emptyset$$

todo camino que une a C y C' contiene a S .

3. Cada distribución condicionada de la red $f_i(x^{\downarrow i}, x^{\downarrow \pi(i)})$ se asigna a un nodo que contenga a la familia formada por X_i y $X_{\pi(i)}$.

Cada nodo C_i lleva asociado un potencial, ϕ_i , éste se calculará como el producto de las distribuciones que hemos asignado a C_i :

$$\phi_{C_i}(x^{\downarrow C_i}) = \prod_{\pi(i) \cup \{i\} \subset C_i} f_i(x^{\downarrow i}, x^{\downarrow \pi(i)})$$

En caso de no tener ninguna distribución asociada se considera $\phi_{C_i} = 1$.

En cuanto a la d-separación en la red bayesiana, ésta queda representada en el árbol de grupos mediante los separadores : "Dos grupos cualesquiera de un árbol de grupos, C y C' , son condicionalmente independientes dado cualquier separador correspondiente a cualquier arco que pertenezca al camino entre C y C' ".

El algoritmo Hugin trabaja sobre un tipo especial de árboles de grupos, llamados árboles de cliques, éstos son en los que ningún nodo es subconjunto de un nodo vecino.

Una vez construido el árbol de cliques y calculados los potenciales asociados a cada clique se incorpora la evidencia a cada uno de los cliques que contenga la variable observada multiplicando el potencial asociado al clique por el potencial asociado a la evidencia, es decir,

$$\delta_{e_i}(x) = \begin{cases} 1 & \text{si } x = e_i \\ 0 & \text{si } x \neq e_i \end{cases}$$

La operación básica del algoritmo HUGIN es la **absorción**

Definición 13. Sean C_i y C_j dos cliques adyacentes con separador S , la operación de absorción es el resultado del proceso siguiente:

$$\begin{aligned}\phi_S^*(x^{\downarrow S}) &= \sum_{x^{\downarrow C_j - S}} \phi_{C_j}(x^{\downarrow C_j}), \quad \forall x \in \Omega_{X_S} \\ \phi_{C_i}^*(x^{\downarrow C_i}) &= \phi_{C_i}(x^{\downarrow C_i}) \cdot \frac{\phi_S^*(x^{\downarrow S})}{\phi_S(x^{\downarrow S})}, \quad \forall x \in \Omega_{X_{C_i}} \\ \phi_{C_j}^*(x^{\downarrow C_j}) &= \phi_{C_j}(x^{\downarrow C_j}), \quad \forall x \in \Omega_{X_{C_j}}\end{aligned}$$

y diremos que C_i ha absorbido de C_j

La idea básica es que la información que C_i y C_j tienen en común se encuentra en el separador S , y es ésta la que C_i recibe de C_j .

Teorema 1. Sea T un árbol de grupos, entonces el producto de los potenciales de los cliques dividido por el producto de los potenciales de los separadores es invariante bajo la absorción.

Demostración. Cuando C_i absorbe de C_j a través de S , sólo se modifican los potenciales de C_i y S , por tanto basta con demostrar que la fracción $\frac{\phi_{C_i}^*}{\phi_S^*}$ es invariante, pero

$$\frac{\phi_{C_i}^*(x^{\downarrow C_i})}{\phi_S^*(x^{\downarrow S})} = \frac{\phi_{C_i}(x^{\downarrow C_i}) \cdot \frac{\phi_S^*(x^{\downarrow S})}{\phi_S(x^{\downarrow S})}}{\phi_S^*(x^{\downarrow S})} = \frac{\phi_{C_i}(x^{\downarrow C_i})}{\phi_S(x^{\downarrow S})}$$

□

Este teorema implica que si, partiendo de una red bayesiana y construyendo su correspondiente árbol de grupos, realizamos una serie de absorciones, T sigue siendo una representación de $P(X)$ y ésta puede calcularse como el cociente entre el producto de los potenciales de los cliques y el producto de los potenciales de los separadores.

Volviendo al proceso de propagación, éste puede verse como un paso de mensajes entre los nodos del árbol en dos fases:

1. **Recolección:** supongamos que C_i recibe un mensaje de recolección de C_j , entonces C_i manda un mensaje de recolección a todos sus vecinos excepto C_j ; cuando un clique termina la operación de recolección, el clique que la solicitó absorbe de éste.
2. **Distribución:** si C_i recibe un mensaje de distribución de C_j , entonces C_i absorbe de C_j y manda un mensaje de distribución a todos sus vecinos salvo C_j .

Teniendo en cuenta ambas fases, el algoritmo de propagación *Hugin* procede de la siguiente manera:

1. Construir el árbol de cliques.
2. Calcular los potenciales asociados a cada clique y separador.
3. Incorporar las evidencias.
4. Seleccionar un clique como raíz.
5. Enviar un mensaje de recolección a partir de la raíz.
6. Enviar un mensaje de distribución desde la raíz.

Ejemplo 7. *Veamos cómo actúa el algoritmo Hugin para el árbol de la figura (3.9):*

- *Seleccionamos el clique G como raíz. G envía un mensaje de recolección a E y F .*
- *E envía un mensaje de recolección a A y absorbe de A , con lo que la evidencia e_1 se propaga hasta E .*
- *Análogamente actúa F con B , C y H .*
- *H envía un mensaje de recolección a D y absorbe de D (propagándose e_2 a H).*
- *F absorbe de B , C y H y e_2 se propaga hasta F .*

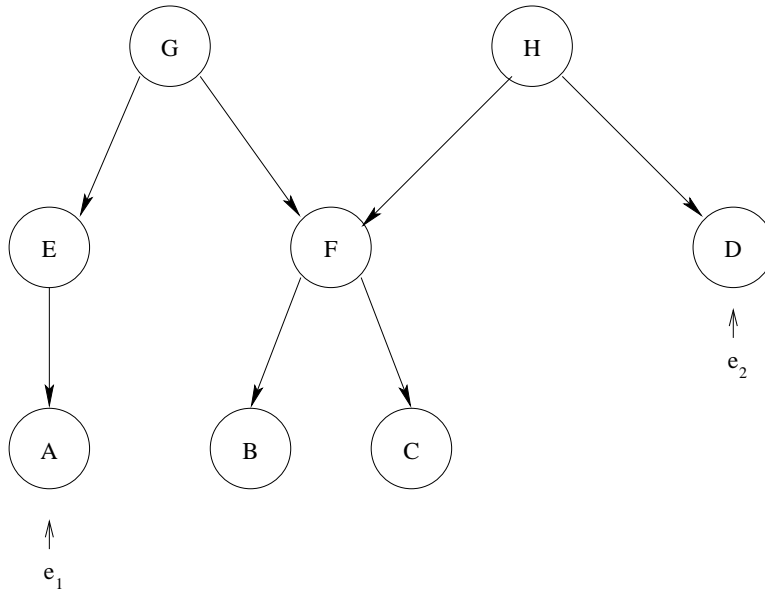


Figura 3.9: Árbol con dos evidencias en los nodos A y D

- En el último paso de esta fase, G absorbe de E y de F con lo que las evidencias e_1 y e_2 han llegado hasta G.
- En la siguiente fase G envía un mensaje de distribución a E y F, que absorben de G ambas evidencias. Cuando E y F envían el mensaje de distribución a todos sus vecinos salvo G, e_1 y e_2 se propagan por toda la red.

Después de la propagación, el árbol resultante es invariante para absorciones, y se cumple que la distribución a posteriori de cualquier variable se puede obtener a partir del potencial de cualquier clique que la contenga mediante marginalización.

3.4. Extracción de perfiles

Una tarea relacionada con la propagación de probabilidades que se puede llevar a cabo sobre redes bayesianas, y que las hace especialmente interesantes para el análisis de datos, es la llamada *inferencia abductiva* o la búsqueda de las explicaciones más

probables, que más adelante definiremos formalmente. En términos de análisis de datos, podemos entender la inferencia abductiva como la búsqueda del perfil más probable de los individuos de una población, bajo determinadas condiciones impuestas por las variables observadas. Este perfil puede venir dado por algunas o todas las variables del sistema (abducción parcial o abducción total, respectivamente).

3.4.1. Abducción total

Definiremos en este apartado los conceptos de explicación y explicación más probable, y veremos cómo se pueden obtener usando el algoritmo HUGIN.

Definición 14. Sea $G = (\Upsilon, \varepsilon)$ una red bayesiana y x_O una observación del conjunto de variables $X_O \subset \Upsilon$. Se dice que $x \in \Omega_X$ es una explicación de x_O si $x \downarrow^{X_O} = x_O$.

Ante una observación x_O , está claro que pueden darse muchas explicaciones, pero quizás sólo algunas de ellas sean de interés. Una manera de elegir entre ellas puede ser en función de su probabilidad ‘a posteriori’.

Definición 15. Sea $G = (\Upsilon, \varepsilon)$ una red bayesiana y x_O una observación del conjunto de variables $X_O \subset \Upsilon$. Se dice que $x \in \Omega_X$ es la explicación más probable (EMP) de x_O si

$$x = \arg \max_{\Omega_X} P(\Upsilon | x_O) . \quad (3.3)$$

Obsérvese que la búsqueda de la explicación más probable no es lo mismo que obtener la configuración formada por los valores de cada variable independientemente que tienen mayor probabilidad a posteriori, pues esto sólo sería válido, en general, en el caso de que todas las variables fueran independientes dos a dos.

Una forma de extraer la explicación más probable [19] consiste en aplicar el algoritmo HUGIN pero sustituyendo el operador de marginalización (suma) por el del máximo; es decir, cuando se aplica la marginalización sobre un separador, en lugar de tomar la suma de los valores de probabilidad para los distintos valores de la variables que se está borrando, se toma el máximo valor de probabilidad. A continuación, una

vez que se ha hecho un paso de mensajes sobre el árbol de cliques, la configuración más probable viene determinada por la unión de las configuraciones de máxima probabilidad en cada clique.

3.4.2. Abducción parcial

En muchos casos, especialmente cuando el número de variables del problema es muy alto, estaremos interesados en perfiles en cuanto a un número reducido de características o variables. En ese caso, hablaremos de abducción parcial, que se puede definir formalmente como sigue.

Definición 16. Sea $G = (\Upsilon, \varepsilon)$ una red bayesiana y x_O una observación del conjunto de variables $X_O \subset \Upsilon$. Sea $X_E \subset \Upsilon$ el conjunto de variables de interés o conjunto explicación. Sea $X_R = \Upsilon \setminus X_E$. Decimos que $x_E \in \Omega_{X_E}$ es la explicación más probable (EMP), en términos de X_E , de x_O , si

$$x = \arg \max_{\Omega_{X_E}} \sum_{X_R} P(X_E, X_R | x_O) . \quad (3.4)$$

3.5. Programa Elvira

El programa Elvira [15], cuyo nombre oficial es *Entorno de Desarrollo para Modelos Gráficos Probabilísticos*, fue desarrollado entre los años 1997 y 2000, como objetivo de un Proyecto Coordinado de I+D, financiado por la CICYT. En este proyecto participaron 25 profesores de 8 universidades españolas agrupadas en cuatro subproyectos: Granada, Almería, País Vasco y UNED.

El programa presenta tres modos básicos:

- Edición
- Inferencia
- Aprendizaje

Introducción a las redes bayesianas

El modo *Aprendizaje* toma la base de datos y construye las tablas de probabilidad y la estructura de la red bayesiana.

Para poder ver mejor la estructura gráfica de la red, utilizamos el modo *Edición*, modificando la red de forma que puedan verse todos los nodos y las relaciones entre ellos.

Una vez creada la red, utilizamos el modo *Inferencia* para propagar probabilidades y extraer perfiles.

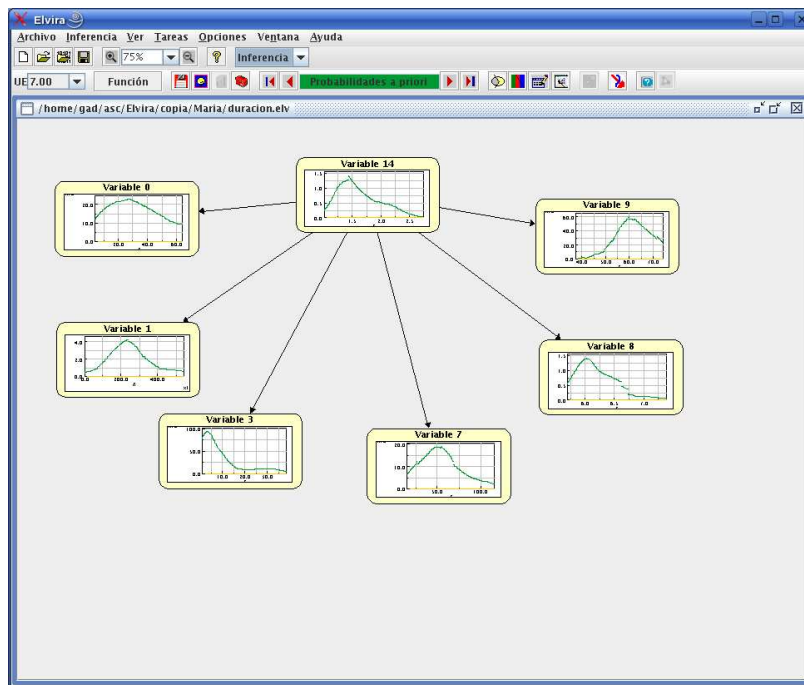


Figura 3.10: Programa Elvira.

Capítulo 4

ANÁLISIS DE INDICADORES DE RENDIMIENTO. EL CASO DE LA UNIVERSIDAD DE ALMERÍA

En este capítulo analizaremos los datos de las asignaturas y titulaciones descritos en el capítulo 2 mediante el uso de redes bayesianas que describan la estructura de dependencias entre las variables.

Dividiremos el capítulo en tres partes, dependiendo del objetivo de nuestros análisis. En la primera parte, el objetivo será el estudio del rendimiento académico de las asignaturas, centrándonos en los factores que influyen en dicho rendimiento y cuáles de ellos la Universidad puede controlar para mejorar los resultados en las asignaturas.

En la segunda realizaremos un estudio análogo pero con los resultados globales obtenidos a lo largo de cuatro cursos en los distintos planes de estudio.

Para finalizar, extraeremos los perfiles de las asignaturas y titulaciones con mejores resultados.

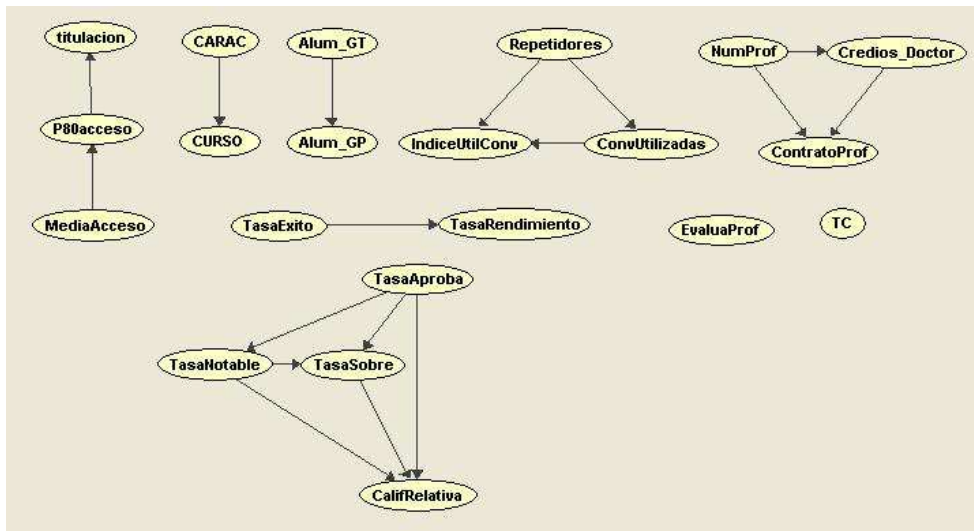


Figura 4.1: Red de asignaturas. Discretización por igual frecuencia y aprendizaje con PC.

4.1. Análisis por asignaturas.

Partiendo de una base de datos con 1.345 registros y 20 características de las asignaturas impartidas en la Universidad de Almería en el curso 2003-04, hemos discretizado las variables continuas para el aprendizaje de la red utilizando intervalos de igual frecuencia e intervalos construidos a través del algoritmo de clustering de K-medias. Para cada base de datos se han aprendido dos redes, utilizando para ello el algoritmo PC al 95% de confianza y el algoritmo K2 tomando como máximo 5 padres. Las redes obtenidas en estos procesos pueden verse en las figuras 4.1, 4.2, 4.3 y 4.4.

Para decidir qué red se ajusta mejor a los datos, hemos calculado el valor del logaritmo de la verosimilitud de la base de datos dada cada una de las cuatro redes, cuyos valores están recogidos en la tabla 4.1.

De estos valores se deduce que la discretización utilizando el algoritmo de K-medias lleva a redes que modelizan mejor que las construidas sobre discretizaciones basadas en intervalos de igual frecuencia. Entre los métodos de aprendizaje, para este caso escogeremos el K2, que es el que crea una red con mayor verosimilitud.

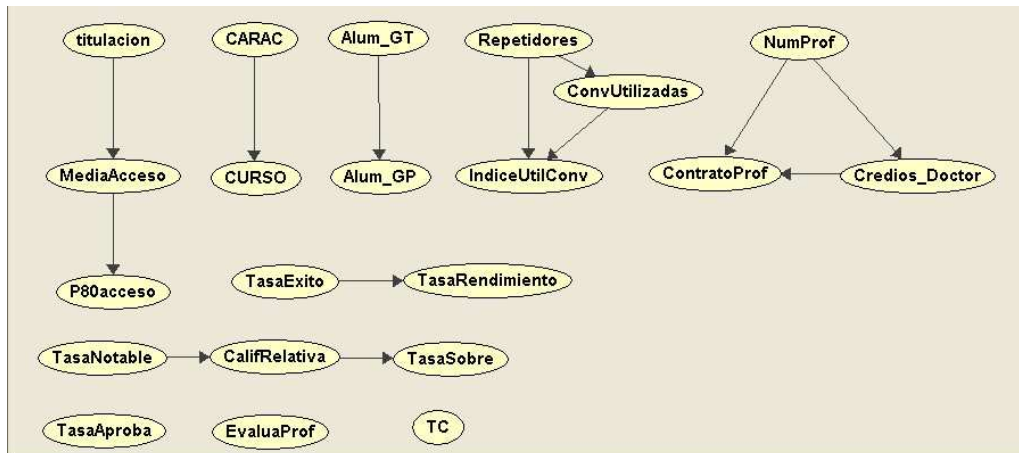


Figura 4.2: Red de asignaturas. Discretización por K-medias y aprendizaje con PC.

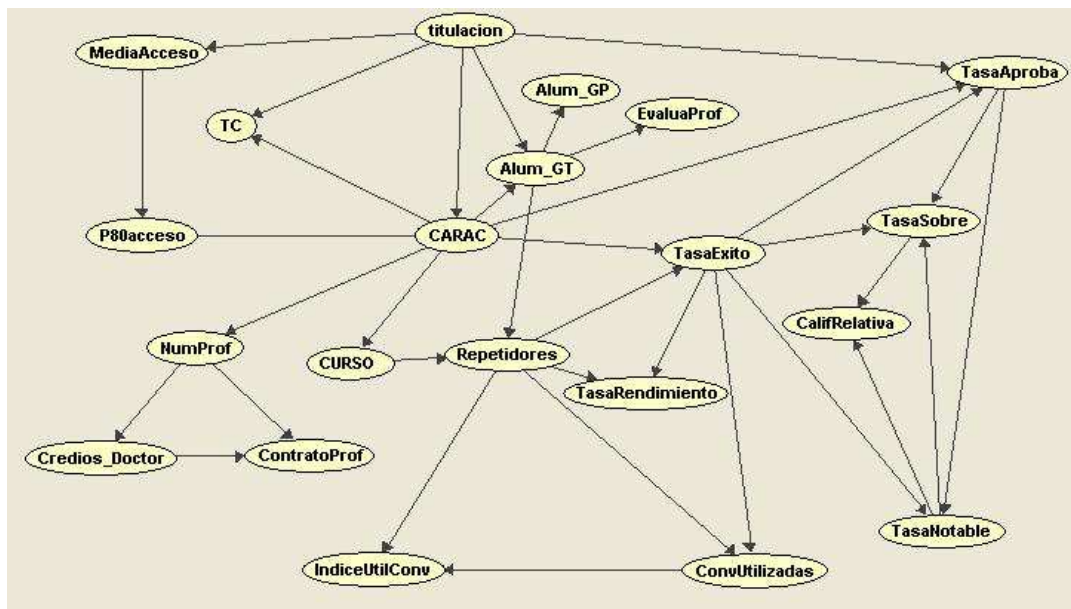


Figura 4.3: Red de asignaturas. Discretización por igual frecuencia y aprendizaje con K2.

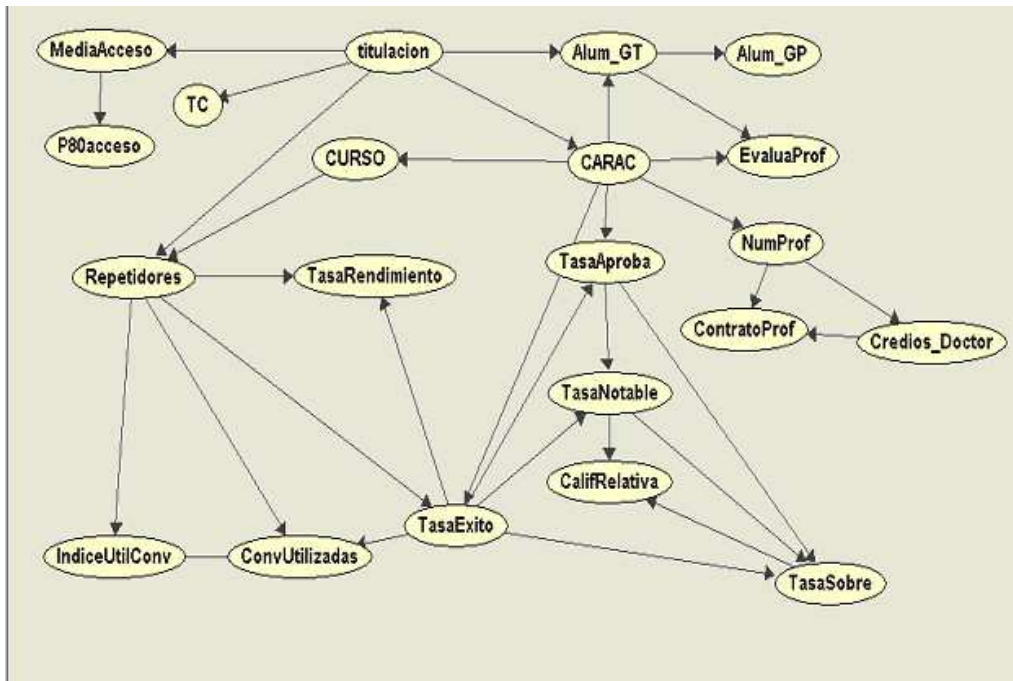


Figura 4.4: Red de asignaturas. Discretización por K-medias y aprendizaje con K2.

DISCRETIZACIÓN	APRENDIZAJE	LOGVERO
Igual frecuencia	PC 95 %	-35.563,1278
K-medias	PC 95 %	-35.071,9166
Igual frecuencia	K2	-31.862,7711
K-medias	K2	-31.143,5305

Tabla 4.1: Logaritmo de la verosimilitud de las redes.

Observando la red de la figura 4.4 concluimos que las variables que influyen sobre los resultados de la asignatura lo hacen a través de las variables *Carácter* y *Titulación*:

- Si se conoce el carácter de la asignatura, el rendimiento de la asignatura es independiente del número de profesores, así como del contrato o la formación de éstos.
- Si se conoce la titulación a la que está adscrita la asignatura, el resultado de ésta es independiente de la nota de acceso o de la dedicación de los alumnos matriculados.
- El número de alumnos por grupo de teoría o práctica, así como la nota obtenida por el profesorado de la asignatura en su evaluación, no influirá en los resultados académicos de la asignatura si se conocen las variables *Carácter* y *Titulación*.

El carácter de la asignatura influye directamente en la tasa de éxito de ésta; para estudiar con más profundidad esta dependencia, propagamos la variable *Carácter* y observamos los cambios en las probabilidades de los valores de la *Tasa de éxito* reflejados en la tabla 4.2.

Los valores de la tabla 4.2 manifiestan claramente que las tasas de éxito altas tienen mayor probabilidad en las asignaturas optativas que en las troncales (la probabilidad de obtener una tasa de éxito superior al 85.5% de los créditos presentados sube de 0.57, en el caso de asignaturas troncales, a 0.84 si se trata de optativas).

Resultados análogos se obtienen si propagamos el *Carácter* de la asignatura y estudiamos las probabilidades de los valores de la *Calificación relativa*, obsérvese en la red 4.4 cómo el carácter de la asignatura es padre de la tasa de aprobados, y, en serie a través de ésta, de las tasas de notable, sobresaliente y de la calificación relativa. Del estudio de los valores de las probabilidades presentados en la tabla 4.3, concluimos que la calificación obtenida por el alumnado en las asignaturas optativas es superior que en las troncales con una mayor probabilidad; obsérvese como la probabilidad de obtener una calificación inferior a 0.315 baja de 0.57 en asignaturas troncales y obligatorias, a 0.39 en optativas.

Tasa de éxito	Probabilidad a Priori	Carácter	
		OBL	OPT
[0, 0.545)	0.03	0.04	0.03
[0.545, 0.735)	0.10	0.15	0.04
[0.735, 0.855)	0.17	0.25	0.09
[0.855, 0.955)	0.20	0.26	0.13
[0.955, 1]	0.49	0.31	0.71

Tabla 4.2: Tasa de éxito en función del carácter de la asignatura.

Calificación Relativa	Probabilidad a Priori	Carácter	
		OBL	OPT
[0, 0.195)	0.23	0.28	0.15
[0.195, 0.315)	0.27	0.29	0.24
[0.315, 0.465)	0.24	0.21	0.28
[0.465, 0.775)	0.18	0.14	0.23
[0.775, 1]	0.08	0.07	0.09

Tabla 4.3: Calif. Relativa en función del carácter de la asignatura.

Para estudiar cómo influye el carácter de la asignatura en el perfil del profesorado que la imparte, propagamos la variable *Character*, obteniendo los resultados recogidos en la tabla 4.4. En ellos no se observa una tendencia clara en cuanto al perfil del profesorado que imparte la asignatura.

Otra conclusión importante que puede extraerse de la red 4.4 es la dependencia directa del tamaño de los grupos de teoría en la evaluación del profesorado. De la tabla 4.5 se deduce una fuerte influencia del nº de alumnos por grupo de teoría en la nota media obtenida por el profesorado de la asignatura, aunque, en términos generales, parece que en las asignaturas con mayor número de alumnos por grupo (siempre y cuando no se superen los 110.75 alumnos por grupo), la probabilidad de notas pequeñas disminuye, aumentando las probabilidades de obtener notas superiores a 3.705. Sin embargo, la probabilidad de obtener una evaluación en los intervalos extremos, es decir, o muy baja (menor de 1.33) o muy alta (superior a 4.285), decrece al aumentar el número de alumnos en el grupo. Este hecho no se debe a un sesgo debido a que la titulación en la que se imparte la asignatura determine el tamaño de los grupos, y sea realmente la titulación la que influye, a través de la variable *Alum_GT*, en la evaluación del profesorado, ya que se ha comprobado que en cada intervalo, están representadas casi todas las titulaciones, o al menos, hay gran variedad de ellas y de todas las ramas de enseñanza.

A continuación pasaremos a estudiar la influencia de la titulación en donde se imparte la asignatura sobre los resultados académicos en ésta.

Para ello, hemos reducido el número de variables de rendimiento académico: se han eliminado las tasas de aprobado, notable y sobresaliente, que quedan representadas por la calificación relativa en la asignatura, y se elimina la tasa de éxito al ser de mayor interés para el estudio la tasa de rendimiento de la asignatura.

Con estas nuevas variables, se sigue el mismo procedimiento para aprender las redes, es decir, se construyen cuatro redes combinando los dos métodos de discretización (intervalos de igual frecuencia y K-medias) y los dos algoritmos de aprendizaje (PC y K2) obteniendo las redes de las figuras 4.5, 4.6, 4.7 y 4.8. Los valores del logaritmo de

Créditos	Probabilidad	Carácter	
		OBL	OPT
Doctor	a Priori		
[0, 0.15)	0.22	0.21	0.24
[0.15, 0.405)	0.04	0.05	0.02
[0.405, 0.63)	0.06	0.08	0.04
[0.63, 0.87)	0.06	0.08	0.04
[0.87, 1]	0.62	0.58	0.66

Créditos	Probabilidad	Carácter	
		OBL	OPT
Numerario	a Priori		
[0, 0.225)	0.24	0.22	0.26
[0.225, 0.52)	0.10	0.13	0.07
[0.52, 0.72)	0.07	0.09	0.04
[0.72, 0.905)	0.05	0.07	0.03
[0.905, 1]	0.54	0.49	0.60

Tabla 4.4: Créditos impartidos por profesorado doctor y funcionario en función del carácter de la asignatura.

Evaluación Profesorado	Probabilidad a Priori	Alumnos por grupo de Teoría				
		< 20.5	[20.5, 44.75)	[44.75, 74.5)	[74.5, 110.75)	≥ 110.75
[0, 1.33)	0.31	0.48	0.32	0.22	0.18	0.23
[1.33, 3.04)	0.09	0.06	0.08	0.12	0.11	0.12
[3.04, 3.705)	0.18	0.06	0.17	0.23	0.25	0.29
[3.705, 4.285)	0.26	0.18	0.25	0.28	0.37	0.24
[4.285, 5]	0.16	0.22	0.17	0.16	0.10	0.12

Tabla 4.5: Evaluación del profesorado en función del nº de alumnos por grupo de teoría.

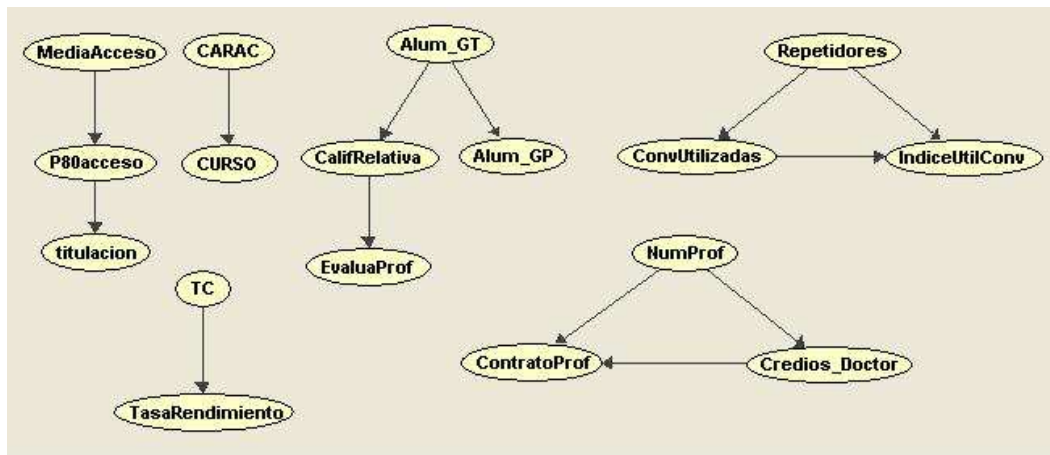


Figura 4.5: Red de asignaturas. Discretización por igual frecuencia y aprendizaje con PC.

la verosimilitud de cada una de las redes, mostrados en la tabla 4.6, determinan que la red que mejor se ajusta a los datos es, de nuevo, la obtenida discretizando a través del K-medias y aprendiendo la red con el algoritmo K2 (figura 4.8).

En esta nueva red, se mantienen las independencias comentadas de la anterior, salvo la de la variable Alum_GT sobre los resultados de la asignatura. Ahora, esta variable influye directamente en la calificación relativa de la asignatura y, a través de ésta, sobre las convocatorias utilizadas, el porcentaje de repetidores y la tasa de rendimiento. La tabla 4.7 muestra los valores de las probabilidades de las variables *CalifRelativa* y

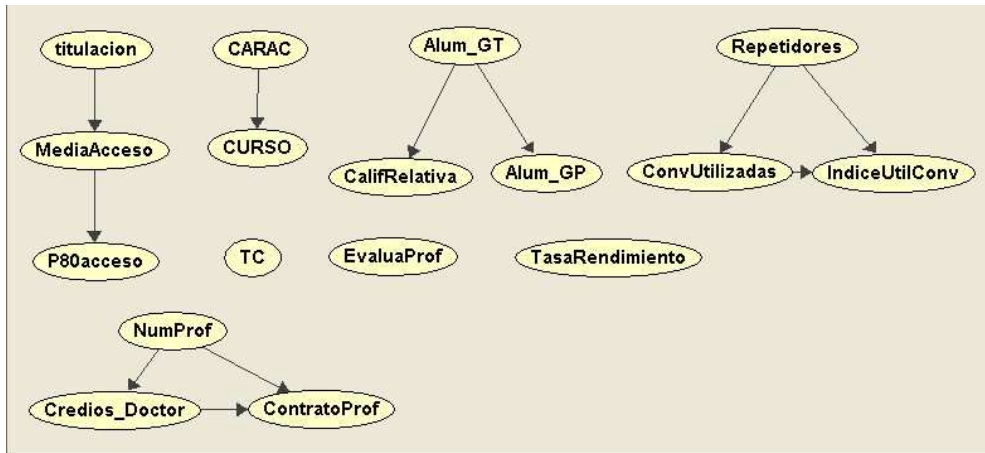


Figura 4.6: Red de asignaturas. Discretización K-medias y aprendizaje con PC.

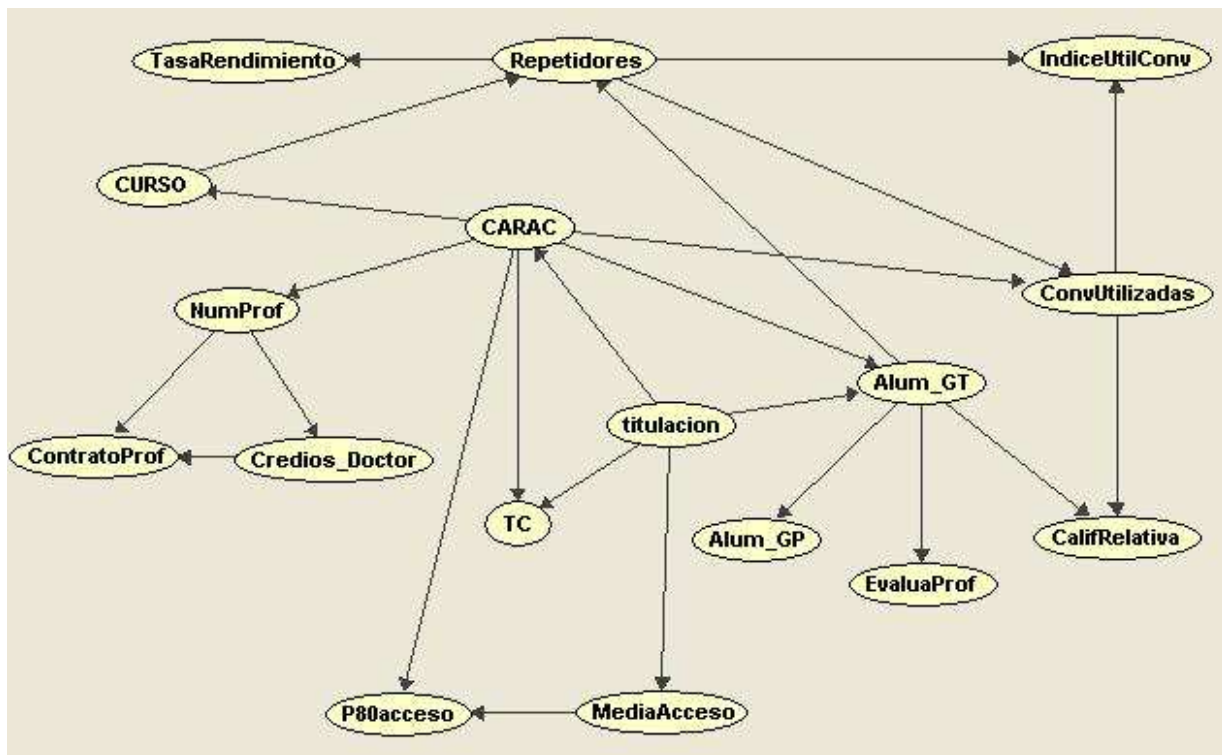


Figura 4.7: Red de asignaturas. Discretización por igual frecuencia y aprendizaje con K2.

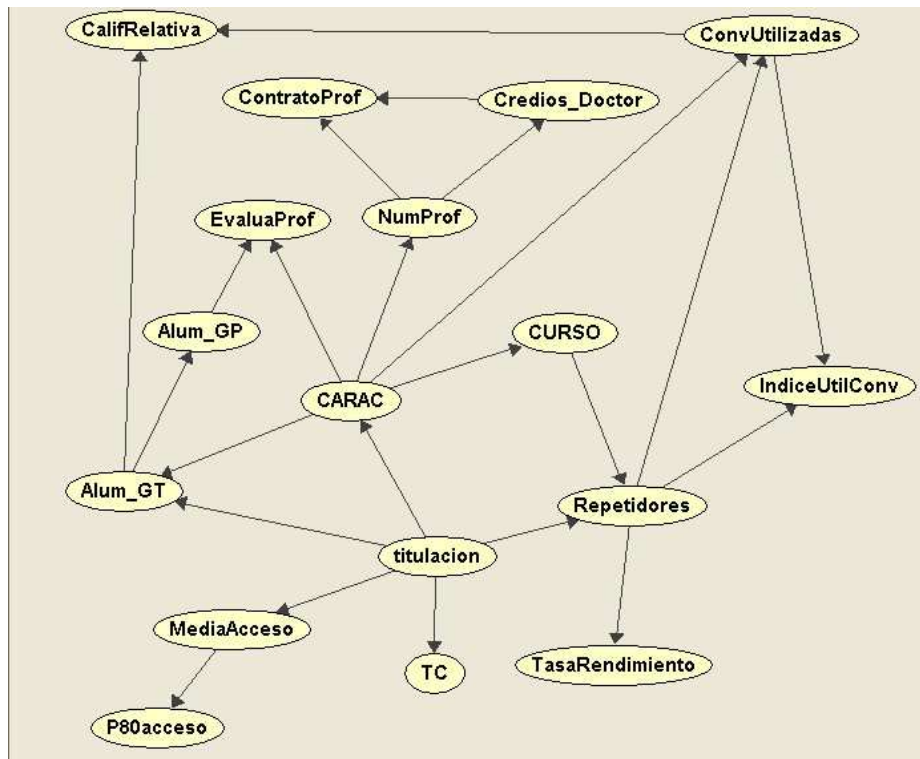


Figura 4.8: Red de asignaturas. Discretización K-medias y aprendizaje con K2.

DISCRETIZACIÓN	APRENDIZAJE	LOGVERO
Igual frecuencia	PC 95 %	-29.156,5705
K-medias	PC 95 %	-28.639,6795
Igual frecuencia	K2	-26.575,9495
K-medias	K2	-25.987,4102

Tabla 4.6: Logaritmo de la verosimilitud de las redes reducido el número de variables.

TasaRendimiento una vez propagada Alum_GT. En ella se puede observar cómo los resultados de las asignaturas dependen fuertemente del tamaño de los grupos de teoría; las probabilidades de obtener calificaciones mayores (superiores a 0.465) disminuyen del 47 % al 14 % conforme aumenta el número de alumnos por grupo, inversamente, mientras que sólo el 22 % de las asignaturas con 20 alumnos o menos por grupo obtienen calificaciones relativas bajas (inferiores a 0.315), este porcentaje sube al 47 % de las asignaturas cuyo alumnado por grupo teórico oscila entre 21 y 44 y sigue aumentando hasta al 73 % de las asignaturas con más de 111 alumnos por grupo teórico.

El comportamiento de la tasa de rendimiento es similar, a mayor número de alumnos por grupo de teoría, menor tasa de rendimiento: la probabilidad de obtener una tasa de rendimiento inferior al 60.5 % de los matriculados sube de 0.42 para grupos con menos de 21 alumnos a 0.52 para grupos con más de 44. Inversamente, la probabilidad de que el porcentaje de alumnos que superan la asignatura sea superior al 75.5 % de los alumnos matriculados baja de 0.35 a 0.26 cuando superamos los 45 alumnos por grupo de teoría. Es interesante el hecho de que las asignaturas con grupos de teoría entre 45 y 74 alumnos presenten las mismas probabilidades de tasas de rendimiento que las asignaturas con grupos teóricos de entre 75 y 110 alumnos y que éstas probabilidades permanezcan casi constantes al superar los 110 alumnos por grupo de teoría, esto nos lleva a que, si quiere aumentarse la tasa de rendimiento de una asignatura, una de las estrategias a seguir es disminuir el número de alumnos por grupo de teoría por debajo de los 44 alumnos.

En la tabla 4.8, se han calculado las probabilidades para el porcentaje de alumnos a tiempo completo matriculados en la asignatura, propagando la variable titulación. En ella queda reflejado cómo el perfil del alumnado cambia según la titulación, pudiéndose hablar de tres tipos de estudios: aquellos con una mayoría de alumnos dedicados a tiempo completo a los estudios (como Matemáticas, Química o Enfermería), estudios con una mayoría de alumnos que compatibilizan los estudios con otras actividades (como Agrónomo, Ingro. en Materiales o Administración y Gestión Pública) y un último grupo de estudios formado por alumnos que se distribuyen más heterogéneamente (Psicopedagogía, Turismo o Magisterio esp. Ed. Primaria). Esto respalda los argumentos del

Calificación Relativa	Probabilidad a Priori	Alumnos por grupo de Teoría				
		< 20.5	[20.5, 44.75)	[44.75, 74.5)	[74.5, 110.75)	≥ 110.75
[0, 0.195)	0.23	0.06	0.16	0.27	0.39	0.46
[0.195, 0.315)	0.29	0.16	0.31	0.37	0.34	0.27
[0.315, 0.465)	0.25	0.32	0.31	0.23	0.16	0.13
[0.465, 0.775)	0.16	0.30	0.17	0.10	0.08	0.11
[0.775, 1]	0.07	0.17	0.05	0.03	0.03	0.03

Tasa de Rendimiento	Probabilidad a Priori	Alumnos por grupo de Teoría				
		< 20.5	[20.5, 44.75)	[44.75, 74.5)	[74.5, 110.75)	≥ 110.75
[0, 0.415)	0.22	0.18	0.21	0.24	0.24	0.25
[0.415, 0.605)	0.27	0.24	0.27	0.28	0.28	0.28
[0.605, 0.755)	0.23	0.23	0.23	0.22	0.22	0.22
[0.755, 0.885)	0.16	0.18	0.16	0.15	0.15	0.14
[0.885, 1]	0.13	0.17	0.13	0.11	0.11	0.12

Tabla 4.7: Resultados académicos en función del nº de alumnos por grupo de teoría.

capítulo 1 sobre *Indicadores*, exponiendo la inconveniencia del uso de indicadores como la *tasa de graduación* o el *coeficiente de duración media de los estudios*, que ignoran la tipología de los alumnos.

Las tablas 4.9 y 4.10 muestran las probabilidades de los porcentajes de alumnos repetidores en la asignatura y las tasas de rendimiento dependiendo de la titulación en la que se imparta. En la primera tabla podemos observar cómo la mayoría de las asignaturas tienen porcentajes de repetidores bajos, inferiores al 19.5%, con mayor probabilidad; sin embargo, cuatro titulaciones se destacan por alcanzar probabilidades altas en los intervalos de mayor porcentaje de repetidores: Derecho y las especialidades de I.T.A., hortofruticultura y jardinería, explotaciones agropecuarias e industrias agrarias. Evidentemente, esto queda respaldado por la tabla 4.10, en la que dichas titulaciones presentan las probabilidades más altas en tasas de aprobados inferiores al 60.5% de los alumnos matriculados, junto con Magisterio especialidad en educación primaria. Las titulaciones con mayor probabilidad de obtener tasas de rendimiento más altas son la especialidad en educación infantil de magisterio, matemáticas y psicopedagogía.

En la red 4.4, estudiamos la fuerte dependencia de la tasa de éxito del carácter de la asignatura. Para la red 4.8, en la que hemos eliminado la tasa de éxito centrando nuestra atención en la tasa de rendimiento, hemos estudiado la variación de las probabilidades de la tasa de rendimiento en función del carácter de la asignatura. Los resultados, recogidos en la tabla 4.11, muestran claramente cómo la tasa de rendimiento en las asignaturas optativas es superior a la de las troncales y obligatorias (la probabilidad de una tasa de rendimiento superior al 88.5% se duplica en el caso de las optativas). Esto corrobora además los resultados de las tablas 4.9 y 4.10 en donde los peores resultados se obtienen en latitulación de Derecho, cuyo plan carece de asignaturas optativas. Por todo esto, con el fin de estudiar las variables que influyen en el rendimiento eliminando las interferencias que el carácter de la asignatura pueda introducir, hemos separado las asignaturas troncales y obligatorias de las optativas, para aprender una red distinta en cada una de estas bases de datos.

<i>TITULACIÓN</i>	Porcentaje de alumnos a tiempo completo				
	< 0.525	[0.525, 0.755)	[0.755, 0.875)	[0.875, 0.955)	≥ 0.955
A Priori	0.05	0.11	0.19	0.35	0.30
Matemáticas	0.04	0.18	0.13	0.22	0.44
Química	0.04	0.04	0.06	0.41	0.45
Agrónomo	0.07	0.50	0.22	0.13	0.09
Derecho	0.07	0.03	0.13	0.70	0.07
Humanidades	0.03	0.30	0.41	0.20	0.06
F.Hispánica	0.02	0.08	0.30	0.28	0.31
Psicología	0.02	0.02	0.06	0.32	0.59
Enfermería	0.03	0.03	0.05	0.16	0.73
Mag.Infantil	0.01	0.01	0.16	0.52	0.29
Mag.Ed.Física	0.03	0.03	0.17	0.53	0.23
Mag.Ed.Primaria	0.04	0.19	0.23	0.35	0.19
Mag.Lengua Ext.	0.03	0.06	0.42	0.39	0.09
Mag.Ed.Musical	0.03	0.10	0.14	0.48	0.24
I.T.A.Horto.	0.04	0.06	0.26	0.52	0.12
I.T.A.Explo.	0.02	0.09	0.34	0.38	0.17
I.T.A.Industrias	0.02	0.02	0.12	0.35	0.49
I.T.A.Mecaniz.	0.04	0.10	0.27	0.39	0.20
Psicopedagogía	0.03	0.36	0.48	0.06	0.06
F.Inglesa	0.01	0.04	0.28	0.43	0.22
I.T.I.Sistemas	0.02	0.02	0.14	0.57	0.24
I.T.I.Gestión	0.02	0.02	0.17	0.59	0.20
Ingro.Sup.Inf.	0.03	0.06	0.26	0.55	0.10
CC.Ambientales	0.02	0.04	0.09	0.54	0.31
I.Químico	0.01	0.03	0.03	0.14	0.78
I.Materiales	0.13	0.67	0.07	0.07	0.07
Empresariales	0.02	0.02	0.14	0.71	0.10
Gestión y A.P.	0.40	0.07	0.02	0.27	0.24
L.A.D.E	0.05	0.14	0.17	0.21	0.44
Turismo	0.02	0.20	0.22	0.20	0.36
CC.Trabajo	0.02	0.02	0.02	0.38	0.55

Tabla 4.8: Porcentaje de alumnos a tiempo completo en función de la titulación.

<i>TITULACIÓN</i>	Porcentaje de repetidores				
	< 0.075	[0.075, 0.195)	[0.195, 0.325)	[0.325, 0.475)	≥ 0.475
A Priori	0.31	0.20	0.17	0.16	0.16
Matemáticas	0.49	0.11	0.13	0.17	0.10
Química	0.32	0.15	0.17	0.14	0.20
Agrónomo	0.32	0.22	0.16	0.15	0.15
Derecho	0.11	0.11	0.16	0.28	0.35
Humanidades	0.20	0.35	0.24	0.09	0.13
F.Hispánica	0.22	0.31	0.18	0.16	0.13
Psicología	0.28	0.20	0.19	0.22	0.12
Enfermería	0.38	0.21	0.16	0.14	0.11
Mag.Infantil	0.46	0.19	0.13	0.11	0.10
Mag.Ed.Física	0.12	0.31	0.24	0.20	0.12
Mag.Ed.Primaria	0.14	0.21	0.17	0.31	0.16
Mag.Lengua Ext.	0.12	0.22	0.28	0.19	0.19
Mag.Ed.Musical	0.19	0.30	0.18	0.21	0.12
I.T.A.Horto.	0.27	0.12	0.19	0.13	0.29
I.T.A.Explo.	0.27	0.12	0.19	0.13	0.29
I.T.A.Industrias	0.26	0.15	0.14	0.17	0.28
I.T.A.Mecaniz.	0.26	0.18	0.14	0.21	0.20
Psicopedagogía	0.27	0.24	0.17	0.16	0.16
F.Inglesa	0.25	0.25	0.23	0.15	0.12
I.T.I.Sistemas	0.29	0.18	0.13	0.20	0.20
I.T.I.Gestión	0.29	0.20	0.15	0.20	0.20
Ingro.Sup.Inf.	0.27	0.23	0.18	0.17	0.15
CC.Ambientales	0.38	0.11	0.22	0.14	0.14
I.Químico	0.38	0.11	0.22	0.14	0.14
I.Materiales	0.29	0.18	0.18	0.18	0.18
Empresariales	0.26	0.19	0.17	0.16	0.22
Gestión y A.P.	0.38	0.19	0.20	0.13	0.10
L.A.D.E	0.43	0.13	0.14	0.16	0.14
Turismo	0.39	0.27	0.12	0.12	0.10
CC.Trabajo	0.33	0.20	0.16	0.16	0.16

Tabla 4.9: Porcentaje de repetidores en la asignatura en función de la titulación.

<i>TITULACIÓN</i>	Tasa de Rendimiento				
	[0, 0.415)	[0.415, 0.605)	[0.605, 0.755)	[0.755, 0.885)	[0.885, 1]
A Priori	0.22	0.27	0.23	0.16	0.13
Matemáticas	0.19	0.24	0.22	0.17	0.18
Química	0.24	0.27	0.21	0.15	0.13
Agrónomo	0.21	0.26	0.23	0.17	0.14
Derecho	0.37	0.32	0.17	0.08	0.06
Humanidades	0.18	0.27	0.25	0.18	0.11
F.Hispánica	0.20	0.27	0.24	0.17	0.11
Psicología	0.21	0.28	0.23	0.15	0.12
Enfermería	0.18	0.25	0.24	0.18	0.15
Mag.Infantil	0.17	0.23	0.23	0.19	0.18
Mag.Ed.Física	0.21	0.30	0.25	0.15	0.08
Mag.Ed.Primaria	0.37	0.21	0.32	0.13	0.08
Mag.Lengua Ext.	0.25	0.31	0.23	0.13	0.07
Mag.Ed.Musical	0.22	0.29	0.24	0.16	0.10
I.T.A.Horto.	0.28	0.28	0.20	0.13	0.11
I.T.A.Explo.	0.28	0.29	0.21	0.13	0.10
I.T.A.Industrias	0.29	0.28	0.19	0.13	0.11
I.T.A.Mecaniz.	0.26	0.28	0.21	0.14	0.11
Psicopedagogía	0.22	0.27	0.23	0.16	0.18
F.Inglesa	0.20	0.28	0.24	0.16	0.12
I.T.I.Sistemas	0.25	0.27	0.21	0.15	0.12
I.T.I.Gestión	0.26	0.28	0.21	0.14	0.11
Ingro.Sup.Inf.	0.22	0.27	0.23	0.16	0.12
CC.Ambientales	0.20	0.27	0.23	0.16	0.14
I.Químico	0.20	0.26	0.23	0.16	0.15
I.Materiales	0.23	0.27	0.22	0.15	0.12
Empresariales	0.25	0.27	0.21	0.14	0.11
Gestión y A.P.	0.18	0.25	0.24	0.17	0.15
L.A.D.E	0.20	0.25	0.22	0.16	0.16
Turismo	0.17	0.24	0.24	0.19	0.16
CC.Trabajo	0.21	0.26	0.22	0.16	0.14

Tabla 4.10: Tasa de rendimiento en las asignaturas en función de la titulación.

Tasa de rendimiento	Probabilidad a Priori	Carácter	
		OBL	OPT
[0, 0.415)	0.22	0.27	0.16
[0.415, 0.605)	0.27	0.30	0.23
[0.605, 0.755)	0.23	0.22	0.24
[0.755, 0.885)	0.16	0.13	0.20
[0.885, 1]	0.13	0.09	0.18

Tabla 4.11: Tasa de rendimiento en función del carácter de la asignatura.

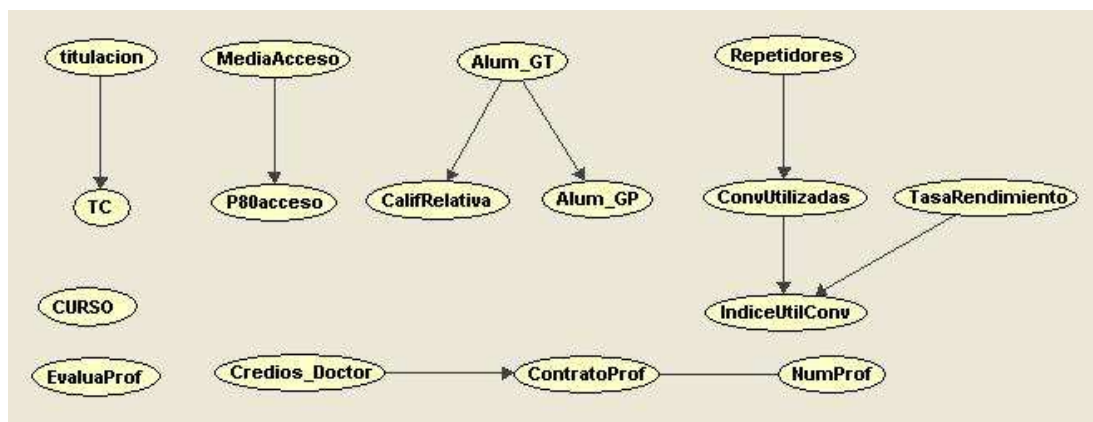


Figura 4.9: Red de asignaturas troncales y obligatorias. Discretización por igual frecuencia y aprendizaje con PC.

4.1.1. Estudio de las asignaturas troncales y obligatorias.

Continuando con el criterio seguido para construir las redes generales, hemos discretizado las variables de las 732 asignaturas a través de los métodos de igual frecuencia y el K-medias, y para cada caso, hemos aprendido dos redes, con el método PC al 95% de confianza y con el K2 con 5 padres máximo. Las redes obtenidas fueron las mostradas en las figuras 4.9, 4.10, 4.11 y 4.12. Los resultados del valor del logaritmo de la verosimilitud, a partir de los cuales determinamos que la red que mejor se ajusta a los datos es la red 4.12, discretizada mediante el algoritmo de K-medias y utilizando el K2 para su aprendizaje, se recogen en la tabla 4.12.

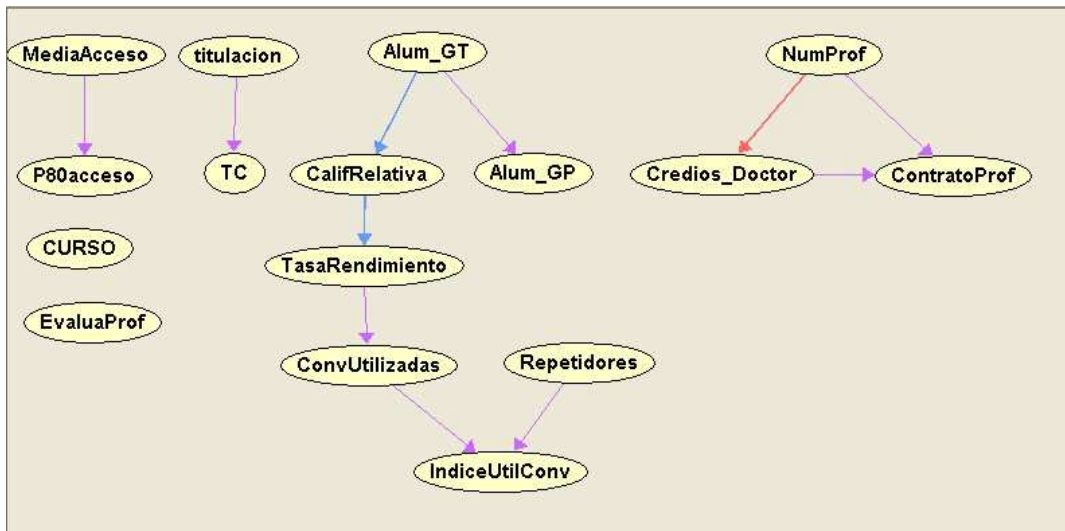


Figura 4.10: Red de asignaturas troncales y obligatorias. Discretización por K-medias y aprendizaje con PC.

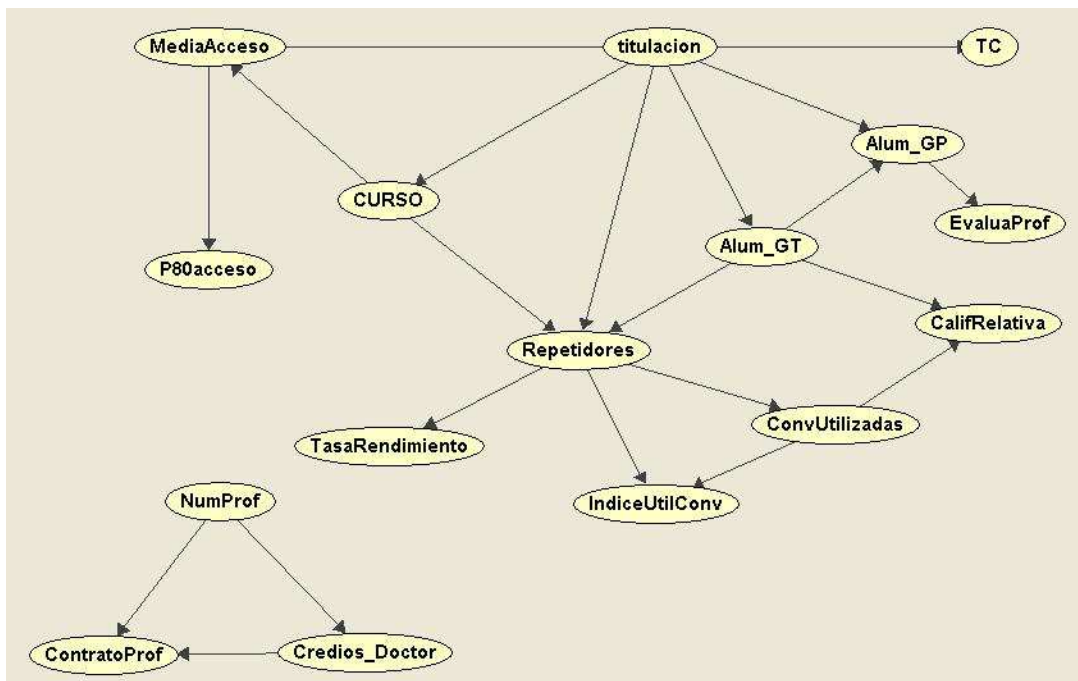


Figura 4.11: Red de asignaturas troncales y obligatorias. Discretización por igual frecuencia y aprendizaje con K2.

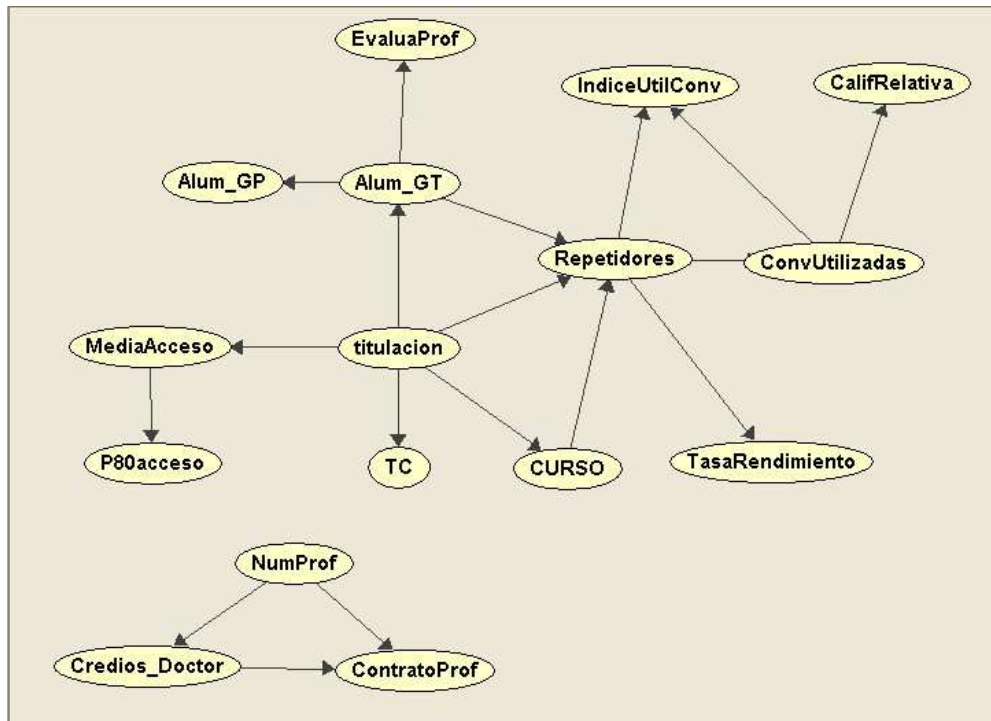


Figura 4.12: Red de asignaturas troncales y obligatorias. Discretización por K-medias y aprendizaje con K2.

DISCRETIZACIÓN	APRENDIZAJE	LOGVERO
Igual frecuencia	PC 95 %	-17.022,1007
K-medias	PC 95 %	-16.308,6508
Igual frecuencia	K2	-14.780,9119
K-medias	K2	-14.346,8951

Tabla 4.12: Logaritmo de la verosimilitud de las redes de asignaturas obligatorias.

Calificación Relativa	Probabilidad a Priori	Alumnos por grupo de teoría				
		< 25.5	[25.5, 49.5)	[49.5, 79.25)	[79.25, 114.75)	≥ 114.75
[0, 0.165)	0.26	0.06	0.18	0.24	0.35	0.63
[0.165, 0.265)	0.33	0.22	0.27	0.39	0.41	0.27
[0.265, 0.405)	0.26	0.33	0.37	0.27	0.17	0.07
[0.405, 0.705)	0.11	0.24	0.13	0.10	0.05	0.02
[0.705, 1]	0.04	0.15	0.04	0.01	0.02	0.01

Tasa de Rendimiento	Probabilidad a Priori	Alumnos por grupo de teoría				
		< 25.5	[25.5, 49.5)	[49.5, 79.25)	[79.25, 114.75)	≥ 114.75
[0, 0.355)	0.19	0.17	0.18	0.18	0.21	0.26
[0.355, 0.535)	0.27	0.21	0.25	0.28	0.30	0.33
[0.535, 0.695)	0.23	0.24	0.24	0.23	0.23	0.22
[0.695, 0.845)	0.18	0.20	0.19	0.19	0.18	0.14
[0.845, 1]	0.11	0.18	0.14	0.11	0.08	0.06

Tabla 4.13: Asignaturas obligatorias: rendimiento en función del alumnado por grupo teórico en la red 4.10.

Cabe destacar cómo, al estudiar sólo las asignaturas troncales y obligatorias, el algoritmo PC al 95 % de confianza encuentra dependencias entre el número de alumnos por grupo de teoría y los resultados de la asignatura. En la red 4.10, gracias al color de los enlaces, podemos ver una dependencia decreciente entre las citadas variables. En la tabla 4.13, aparecen las probabilidades de los valores de la calificación relativa y de la tasa de rendimiento de las asignaturas al propagar, en la red 4.10, el número de alumnos por grupo de teoría. Esta tabla pone en evidencia el aumento del rendimiento al disminuir el número de alumnos por grupo de teoría: mientras que el 63 % de las asignaturas troncales y obligatorias con más de 115 alumnos por grupo teórico obtienen notas relativas muy bajas, este porcentaje se reduce al 6 % de las asignaturas con menos de 25 alumnos por grupo de teoría. En cuanto al intervalo más alto de calificaciones, se obtienen en el 15 % de las asignaturas con un máximo de 25 alumnos por grupo, bajando este porcentaje hasta al 1 % al ir aumentando el tamaño de los grupos teóricos.

Un comportamiento equivalente presenta la tasa de rendimiento al propagar la variable *Alumn_GT*: las probabilidades de tasas de rendimiento inferiores al 53.5 % suben del 38 % de las asignaturas con menor número de alumnos por grupo de teoría al 59 % de las asignaturas con más de 115 alumnos por grupo, cayendo las probabilidades de tasas de rendimiento altas de 0.38 a 0.20.

En este apartado nos centraremos en el estudio de la red 4.12 y de las dependencias que refleja.

Cuando estudiamos todas las asignaturas juntas, el perfil del profesorado (número de profesores que imparten la asignatura y si son doctores o permanentes) influía en los resultados de la asignatura a través del carácter de ésta. Una vez eliminada la variable carácter, la red 4.12 muestra cómo el rendimiento de la asignatura es independiente del perfil del profesorado que la imparte.

Las variables titulación, número de alumnos por grupo de teoría y curso son las únicas que influyen directamente sobre los resultados de la asignatura y, a través de ellas, el resto de variables, dándose las siguientes independencias condicionadas:

- El número de alumnos por grupo de prácticas, sólo influye en los resultados si no se conoce el número de alumnos por grupo de teoría. Lo mismo ocurre con la variable que recoge la evaluación de los docentes de la asignatura.
- El porcentaje de alumnos a tiempo completo así como la nota de acceso de los alumnos matriculados en la asignatura tendrán un peso sobre sus resultados siempre que se desconozca la titulación de la asignatura.

Antes de comenzar con el estudio de la dependencia de los resultados de la asignatura, es interesante estudiar cómo el número de alumnos por grupo de teoría influye en la evaluación obtenida por el profesorado de la asignatura. De la tabla 4.14 pueden obtenerse tres conclusiones principales:

- Las probabilidades de una evaluación muy negativa (inferior a 1.33 puntos) son mayores cuanto más reducido es el grupo de teoría. Esto puede deberse a que hemos puntuado con cero a las asignaturas no evaluadas. Estudiando los datos, de las 156 asignaturas no evaluadas, 43 tienen menos de 26 alumnos por grupo, 37 entre 26 y 49, 43 entre 50 y 79 alumnos y 23 y 10 asignaturas pertenecientes a los últimos dos intervalos.

Evaluación del Profesorado	Probabilidad a Priori	Alumnos por grupo de teoría				
		< 25.5	[25.5, 49.5)	[49.5, 79.25)	[79.25, 114.75)	≥ 114.75
< 1.33	0.216	0.39	0.22	0.19	0.16	0.13
[1.33, 3.01)	0.12	0.10	0.11	0.13	0.11	0.12
[3.01, 3.685)	0.24	0.09	0.20	0.26	0.26	0.43
[3.685, 4.245)	0.28	0.18	0.29	0.28	0.36	0.27
[4.245, 5]	0.15	0.24	0.18	0.14	0.11	0.06

Tabla 4.14: Asignaturas obligatorias: evaluación del profesorado en función del nº de alumnos por grupo teórico.

- Si nos centramos en las probabilidades de una evaluación superior a 1.33 puntos, apenas hay variación entre las probabilidades de una evaluación deficiente (menor de 3 puntos) al aumentar el número de alumnos por grupo.
- Mientras que las mayores puntuaciones se obtienen en asignaturas con menos de 25 alumnos por grupo de teoría, ésta probabilidad va bajando al aumentar el número de alumnos, desplazándose los valores más probables a evaluaciones medias (entre 3.01 y 3.685 puntos)

Por lo tanto se puede concluir que la consideración del profesor por parte de los alumnos mejora cuanto menores son los grupos de teoría en la asignatura.

El resto del epígrafe lo dedicaremos al estudio de los efectos de las distintas variables consideradas en la red 4.12 sobre la tasa de rendimiento de las asignaturas.

El número de alumnos por grupo de teoría tiene una relación directa con el porcentaje de repetidores de la asignatura y, a través de ésta, sobre la tasa de rendimiento. La tabla 4.15 muestra claramente la fuerte dependencia del porcentaje de alumnos que superan la asignatura con el número de alumnos por grupo de teoría, subiendo la probabilidad de obtener tasas de rendimiento bajas al aumentar el número de alumnos mientras disminuyen las probabilidades de los valores altos de rendimiento. Además, los comportamientos del rendimiento de las clases [49.5, 79.25) y [79.25, 114.75) son idénticos, lo que concluye que para obtener una mejora de la tasa de rendimiento de una asignatura habría que disminuir el número de alumnos por grupo teórico por debajo de 49.

Tasa de Rendimiento	Probabilidad a Priori	Alumnos por grupo de Teoría				
		< 25.5	[25.5, 49.5)	[49.5, 79.25)	[79.25, 114.75)	≥ 114.75
[0, 0.355)	0.20	0.18	0.19	0.20	0.20	0.24
[0.355, 0.535)	0.28	0.26	0.27	0.28	0.28	0.29
[0.535, 0.695)	0.23	0.23	0.23	0.23	0.23	0.22
[0.695, 0.845)	0.18	0.20	0.19	0.18	0.18	0.15
[0.845, 1]	0.11	0.13	0.12	0.11	0.11	0.09

Tabla 4.15: Asignaturas obligatorias: tasa de rendimiento en función del nº de alumnos por grupo de teoría.

Tasa de Rendimiento	Probabilidad a Priori	Alumnos por grupo de Prácticas				
		< 23.9	[23.9, 41.65)	[41.65, 68.5)	[68.5, 114.16)	≥ 114.16
[0, 0.355)	0.20	0.19	0.20	0.21	0.20	0.24
[0.355, 0.535)	0.28	0.27	0.28	0.28	0.28	0.29
[0.535, 0.695)	0.23	0.23	0.23	0.23	0.23	0.24
[0.695, 0.845)	0.18	0.19	0.18	0.18	0.18	0.15
[0.845, 1]	0.11	0.12	0.11	0.11	0.11	0.10

Tabla 4.16: Asignaturas obligatorias: tasa de rendimiento en función del nº de alumnos por grupo de prácticas.

Si se desconoce el valor de la variable *Alumn_GT*, el número de alumnos por grupo de prácticas influye en el porcentaje de repetidores en la asignatura y, a través de ésta, en la tasa de rendimiento. Las probabilidades de la tasa de rendimiento de las asignaturas en función de *Alumn_GP* están recogidas en la tabla 4.16. El efecto de esta variable es menor que el ejercido por los grupos teóricos: las probabilidades de los rendimientos más bajos aumentan con el número de alumnos en el grupo de prácticas, disminuyendo las probabilidades de las tasas más altas. La tasa de rendimiento entre el 53.5% y el 69.5% de los créditos matriculados permanece constante al aumentar el número de alumnos por grupo práctico. En este caso, la distribución de las probabilidades para los intervalos [23.9, 41.65), [41.65, 68.5), [68.5, 114.16) son prácticamente iguales, y muy parecidas a las probabilidades obtenidas para los grupos con menos de 24 alumnos, por lo que variar el tamaño de los grupos prácticos entre estas cifras no conllevará una mejora de la tasa de rendimiento de la asignatura.

El mismo esquema de dependencias sigue la variable *EvaluaProf*, sin embargo el estudio de las probabilidades de la tasa de rendimiento para la propagación de ésta variable (ver tabla 4.17) revela que su efecto sobre el rendimiento en la asignatura no

Tasa de Rendimiento	Probabilidad a Priori	Evaluación del profesorado				
		< 1.33	[1.33, 3.01)	[3.01, 3.685)	[3.685, 4.245)	≥ 4.245
[0, 0.355)	0.20	0.20	0.20	0.21	0.20	0.19
[0.355, 0.535)	0.28	0.27	0.28	0.28	0.28	0.27
[0.535, 0.695)	0.23	0.23	0.23	0.23	0.23	0.23
[0.695, 0.845)	0.18	0.18	0.18	0.17	0.18	0.18
[0.845, 1]	0.11	0.12	0.11	0.11	0.11	0.12

Tabla 4.17: Asignaturas obligatorias: tasa de rendimiento en función de la evaluación del profesorado.

es significativo.

El perfil del alumnado de la asignatura (reflejado en la nota de acceso de los alumnos y el porcentaje de matriculados dedicados a tiempo completo al estudio) influye en los resultados de la asignatura a través de la variable *Titulación*. Como puede deducirse de la tabla 4.18, el efecto que ejercen estas variables sobre la tasa de rendimiento no es importante. Se aprecia una pequeña mejora en la tasa de rendimiento cuando la nota media de acceso del alumnado es superior a 6.46 y el percentil 80 es superior a 7.3. Análogamente, el porcentaje de alumnos a tiempo completo apenas influye en la tasa.

Como conclusión, la variable con un mayor efecto sobre la tasa de rendimiento en asignaturas troncales y obligatorias es el tamaño de los grupos de teoría y, en menor medida, el tamaño de los grupos prácticos, por lo que la actuación más eficaz que podría llevarse a cabo para mejorar el rendimiento en este tipo de asignatura sería disminuir el tamaño de los grupos teóricos por debajo de los 49 alumnos. La nota de acceso del alumnado también tiene un pequeño influjo sobre la tasa de rendimiento, pero apenas significativo.

4.1.2. Estudio de las asignaturas optativas.

Una vez determinadas las variables de las que depende el rendimiento de una asignatura troncal u obligatoria, pasamos a estudiar el caso de las asignaturas optativas. Para ello partimos de una base de datos con 613 asignaturas optativas, y discretizamos las variables en intervalos de igual frecuencia y con el método de K-medias. Para ambos casos aprendemos sendas redes utilizando el algoritmo PC al 95 % de confianza y el K2

Tasa de Rendimiento	Probabilidad a Priori	Nota media de acceso				
		[5.32, 5.92)	[5.92, 6.17)	[6.17, 6.46)	[6.46, 6.83)	[6.83, 7.71]
[0, 0.355)	0.20	0.19	0.21	0.21	0.19	0.19
[0.355, 0.535)	0.28	0.28	0.28	0.28	0.27	0.27
[0.535, 0.695)	0.23	0.24	0.23	0.23	0.23	0.23
[0.695, 0.845)	0.18	0.18	0.17	0.17	0.18	0.19
[0.845, 1]	0.11	0.11	0.11	0.11	0.12	0.13

Tasa de Rendimiento	Probabilidad a Priori	Percentil 80 de la nota de acceso				
		[5.46, 6.30)	[6.30, 6.65)	[6.65, 7.30)	[7.30, 7.61)	[7.61, 8.5]
[0, 0.355)	0.20	0.19	0.21	0.21	0.20	0.19
[0.355, 0.535)	0.28	0.28	0.28	0.28	0.27	0.27
[0.535, 0.695)	0.23	0.23	0.23	0.23	0.23	0.23
[0.695, 0.845)	0.18	0.18	0.18	0.18	0.18	0.19
[0.845, 1]	0.11	0.11	0.11	0.11	0.12	0.12

Tasa de Rendimiento	Probabilidad a Priori	Porcentaje de alumnos a tiempo completo				
		[0.26, 0.55)	[0.55, 0.765)	[0.765, 0.875)	[0.875, 0.945)	≥ 0.945
[0, 0.355)	0.20	0.19	0.19	0.20	0.21	0.20
[0.355, 0.535)	0.28	0.27	0.27	0.28	0.28	0.27
[0.535, 0.695)	0.23	0.23	0.23	0.23	0.23	0.23
[0.695, 0.845)	0.18	0.19	0.19	0.18	0.17	0.18
[0.845, 1]	0.11	0.12	0.12	0.11	0.11	0.12

Tabla 4.18: Asignaturas obligatorias: tasa de rendimiento en función del perfil del alumnado.

con un máximo de 5 padres. Las redes obtenidas se muestran en las figuras 4.13, 4.14, 4.15 y 4.16.

La tabla 4.19 recoge los valores del logaritmo de la verosimilitud de cada una de las cuatro redes, mostrando que, al contrario de los casos anteriores, la red que mejor se ajusta a los datos es la aprendida con el método K2 con una discretización en intervalos de igual frecuencia, aunque la diferencia con la red obtenida sobre una discretización realizada con el algoritmo de K-medias es muy pequeña.

Del análisis de la red 4.15 deducimos que, como en el caso de las asignaturas troncales y obligatorias, el rendimiento en la asignatura es independiente del perfil del profesorado que la imparte, sin embargo, al contrario de las asignaturas de elección obligatoria, la evaluación del profesorado de la asignatura es independiente del resto de variables y, por tanto, tampoco influye en el rendimiento de ésta.

La nota de acceso de los alumnos de la asignatura está conectada a la tasa de

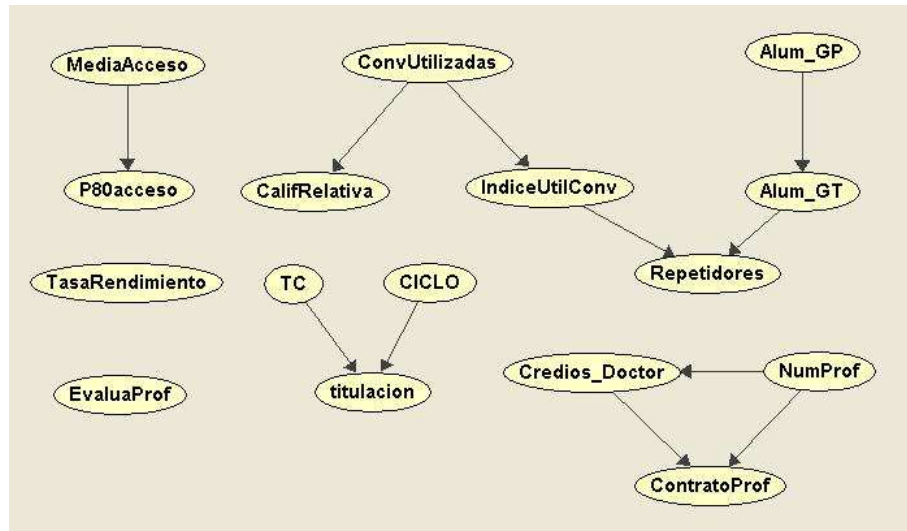


Figura 4.13: Red de asignaturas optativas. Discretización por igual frecuencia y aprendizaje con PC.

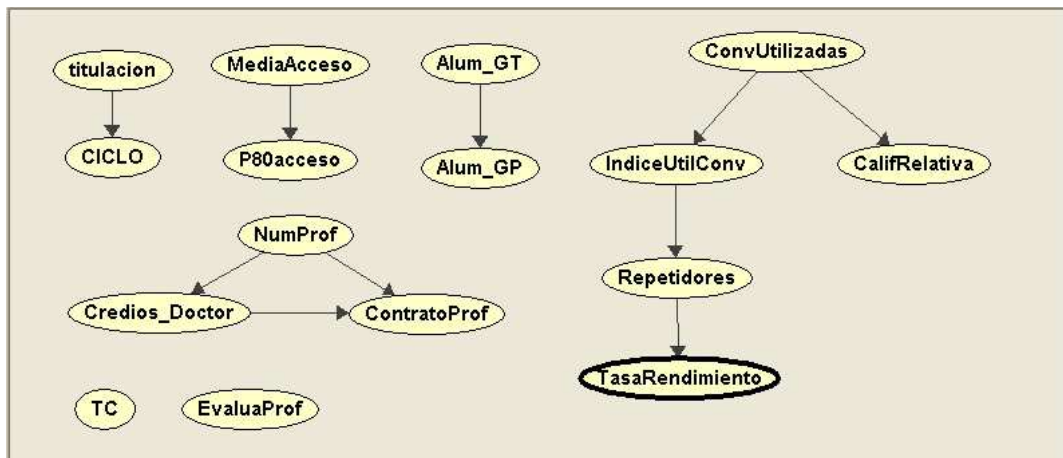


Figura 4.14: Red de asignaturas optativas. Discretización por K-medias y aprendizaje con PC.

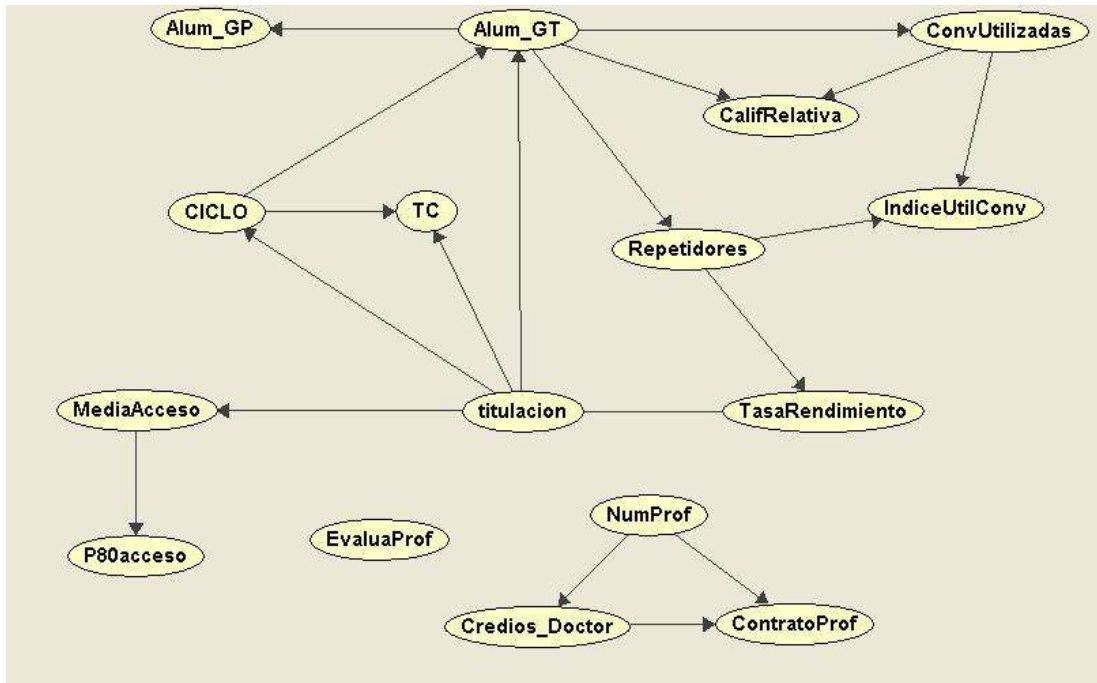


Figura 4.15: Red de asignaturas optativas. Discretización por igual frecuencia y aprendizaje con K2.

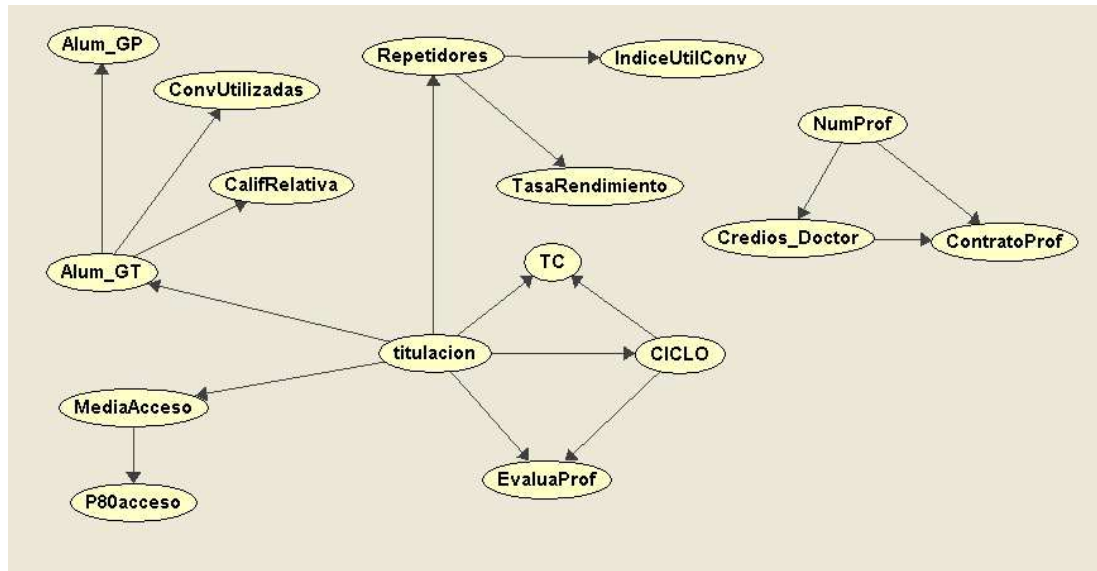


Figura 4.16: Red de asignaturas optativas. Discretización por K-medias y aprendizaje con K2.

DISCRETIZACIÓN	APRENDIZAJE	LOGVERO
Igual frecuencia	PC 95 %	-11.332,9724
K-medias	PC 95 %	-11.408,2034
Igual frecuencia	K2	-10.375,0703
K-medias	K2	-10.490,2214

Tabla 4.19: Logaritmo de la verosimilitud de las redes de asignaturas optativas.

Tasa de Rendimiento	Probabilidad a Priori	Nota media de acceso				
		[5.32, 6.00)	[6.00, 6.20)	[6.20, 6.44)	[6.44, 6.69)	[6.69, 8.03]
[0, 0.59)	0.20	0.18	0.19	0.22	0.22	0.17
[0.59, 0.71)	0.19	0.17	0.20	0.29	0.19	0.18
[0.71, 0.82)	0.20	0.21	0.21	0.21	0.21	0.18
[0.82, 0.92)	0.20	0.23	0.21	0.19	0.19	0.19
[0.92, 1]	0.21	0.20	0.19	0.19	0.20	0.28

Tasa de Rendimiento	Probabilidad a Priori	Percentil 80 de la nota de acceso				
		[5.37, 6.28)	[6.28, 6.61)	[6.61, 6.92)	[6.92, 7.22)	[7.22, 8.57]
[0, 0.59)	0.20	0.19	0.19	0.21	0.21	0.18
[0.59, 0.71)	0.19	0.18	0.19	0.19	0.19	0.18
[0.71, 0.82)	0.20	0.21	0.21	0.21	0.20	0.18
[0.82, 0.92)	0.20	0.22	0.21	0.20	0.19	0.19
[0.92, 1]	0.21	0.20	0.20	0.20	0.21	0.26

Tabla 4.20: Asignaturas optativas: tasa de rendimiento en función de la nota de acceso.

rendimiento a través de la variable titulación, por lo que sólo afectará al rendimiento de la asignatura si la titulación donde se imparte es desconocida. La tabla 4.20 muestra las probabilidades de la tasa de rendimiento al propagar una de las dos variables. Dichas probabilidades denotan un mayor rendimiento al aumentar la media de acceso por encima de 6.69 o si el percentil 80 supera los 7.22 puntos.

También la tabla 4.21 revela un efecto importante del porcentaje de alumnos con dedicación completa sobre la tasa de rendimiento.

El tamaño de los grupos de teoría cobra en las asignaturas optativas una mayor relevancia que en el caso de las obligatorias, puesto que es padre de la calificación relativa de la asignatura, del número de convocatorias medio necesario para superarla, y del porcentaje de repetidores, relacionándose a través de ésta con la tasa de rendimiento.

Tasa de Rendimiento	Probabilidad a Priori	Porcentaje de alumnos a tiempo completo				
		[0, 0.77)	[0.77, 0.88)	[0.88, 0.94)	[0.94, 1)	1
[0, 0.59)	0.20	0.19	0.22	0.20	0.20	0.18
[0.59, 0.71)	0.19	0.18	0.19	0.19	0.19	0.18
[0.71, 0.82)	0.20	0.20	0.21	0.21	0.21	0.19
[0.82, 0.92)	0.20	0.19	0.19	0.21	0.22	0.19
[0.92, 1]	0.21	0.23	0.20	0.19	0.19	0.26

Tabla 4.21: Asignaturas optativas: tasa de rendimiento en función del porcentaje de alumnos a tiempo completo.

La tabla 4.22, resume las probabilidades de la calificación y de la tasa de rendimiento de las asignaturas en función del número de alumnos por grupo teórico. En dicha tabla podemos observar el destacado efecto que el tamaño de los grupos tiene tanto sobre la calificación relativa como en la tasa de rendimiento: la probabilidad de calificaciones bajas crece al aumentar el tamaño de los grupos pasando del 16 % en grupos de menos de 10 alumnos al 62 % en grupos de 51 o más alumnos. El 21 % de las asignaturas con menos de 10 alumnos obtuvieron calificaciones superiores a 0.57, mientras que sólo el 7 % de los grupos con un mínimo de 51 alumnos alcanzaron esa nota.

La tasa de rendimiento también está fuertemente relacionada con el tamaño de los grupos de teoría, creciendo las probabilidades de las tasas de rendimiento más bajas al aumentar el número de alumnos y disminuyendo, del 50 % al 39 % la probabilidad de alcanzar tasas de rendimiento superiores al 82 % de los créditos matriculados.

Cualquier actuación que se lleve a cabo para bajar el número de alumnos por grupo de teoría de una clase a otra comporta una mejora significativa de los resultados académicos de la asignatura optativa.

En cuanto a la dependencia entre rendimiento y tamaño de los grupos prácticos, siempre que se desconozca el valor de $Alum_GT$, la tabla 4.23 muestra que las variaciones de probabilidad no son significativas, salvo para el caso de tasas de rendimiento superiores al 92 % de los créditos matriculados que disminuye notablemente al incrementar el número de alumnos: del 30 % de las asignaturas con menos de 9 alumnos o el 23 % de los grupos prácticos entre 9 y 17 alumnos, decrece hasta el 17 % de las asignaturas con más de 38 alumnos por grupo de prácticas.

Calificación Relativa	Probabilidad a Priori	Alumnos por grupo de teoría				
		< 10	[10, 18)	[18, 28)	[28, 51)	[51, 137]
[0, 0.27)	0.18	0.06	0.09	0.13	0.25	0.36
[0.27, 0.37)	0.21	0.10	0.19	0.19	0.30	0.26
[0.37, 0.45)	0.19	0.13	0.24	0.22	0.15	0.18
[0.45, 0.57)	0.21	0.28	0.22	0.22	0.20	0.13
[0.57, 1]	0.21	0.43	0.25	0.23	0.10	0.07

Tasa de Rendimiento	Probabilidad a Priori	Alumnos por grupo de teoría				
		< 10	[10, 18)	[18, 28)	[28, 51)	[51, 137]
[0, 0.59)	0.20	0.18	0.20	0.19	0.21	0.20
[0.59, 0.71)	0.19	0.15	0.19	0.19	0.20	0.19
[0.71, 0.82)	0.20	0.17	0.20	0.21	0.21	0.22
[0.82, 0.92)	0.20	0.19	0.19	0.19	0.21	0.23
[0.92, 1]	0.21	0.31	0.22	0.21	0.17	0.16

Tabla 4.22: Resultados de las asignaturas optativas en función del tamaño del grupo de teoría.

Tasa de Rendimiento	Probabilidad a Priori	Alumnos por grupo práctico				
		< 9	[9, 17)	[17, 25)	[25, 38)	≥ 38
[0, 0.59)	0.20	0.18	0.20	0.20	0.20	0.20
[0.59, 0.71)	0.19	0.16	0.18	0.19	0.19	0.19
[0.71, 0.82)	0.20	0.17	0.20	0.21	0.21	0.21
[0.82, 0.92)	0.20	0.19	0.19	0.20	0.21	0.22
[0.92, 1]	0.21	0.30	0.23	0.20	0.18	0.17

Tabla 4.23: Asignaturas optativas: tasa de rendimiento en función del tamaño de los grupos prácticos.

Como conclusión de este análisis obtenemos que la medida más eficaz para mejorar el rendimiento de las asignaturas optativas es, de nuevo, disminuir el tamaño de los grupos de teoría. La reducción del tamaño de los grupos prácticos no ofrece importantes mejoras en el rendimiento, así como tampoco tiene ningún efecto el profesorado que la imparte.

4.2. Análisis por titulación.

En esta sección, hemos estudiado las titulaciones de la Universidad de Almería en los cursos 2000-01, 2001-02, 2002-03 y 2003-04, observándose distintos planes de una misma titulación en este periodo y sus resultados. El objetivo del análisis es reducir la tasa de abandono y la duración media de los estudios, aumentando las tasa de rendimiento y de graduación. Para ver qué variables pueden influir en estas tasas, hemos escogido seis indicadores de las características del estudio (peso de las prácticas en plan de estudios, flexibilidad curricular a través del estudio de la optatividad y masificación en la titulación), un indicador del tipo de profesorado que imparte clase en la titulación (porcentaje de créditos impartidos por profesores doctores) y dos indicadores del perfil del alumnado de la titulación (nota de acceso y dedicación).

Para efectuar el análisis mediante redes bayesianas de las 102 promociones a lo largo de los cuatro cursos, se ha seguido el mismo procedimiento que en el caso del estudio del rendimiento en las asignaturas: se discretizan las variables continuas mediante intervalos de igual frecuencia y el algoritmo de K-medias, y para cada base de datos, se aprende una red mediante los métodos PC al 95% de confianza y el K2, fijando en cinco el número máximo de padres.

Las redes obtenidas pueden verse en las figuras 4.17, 4.18, 4.19, 4.20.

Como puede verse en la tabla 4.24, que recoge los valores de los logaritmos de las verosimilitudes de las cuatro redes, la red que mejor se ajusta a los datos es la construida con el método K2 sobre variables discretizadas mediante el algoritmo de K-medias.

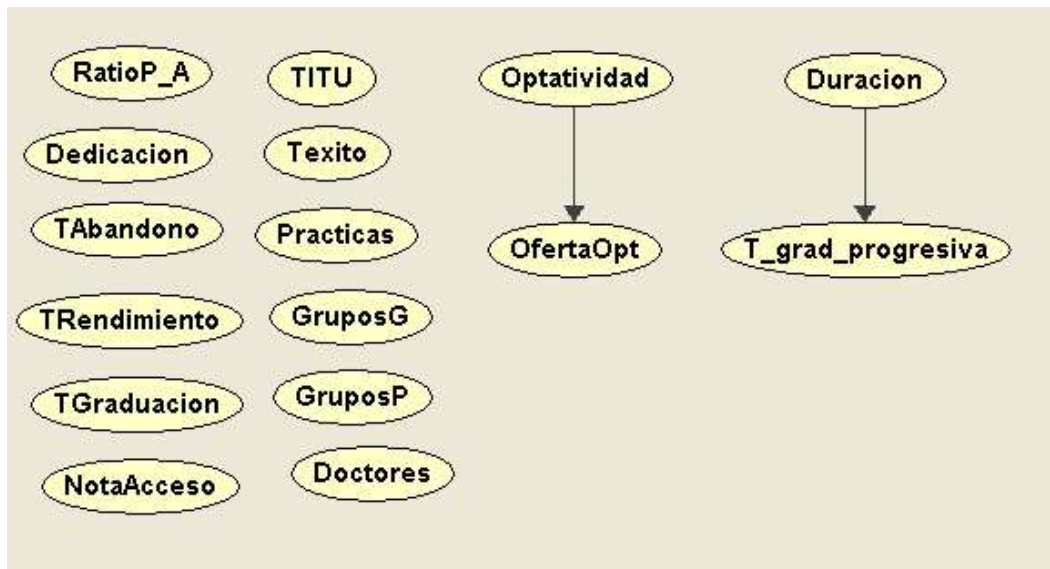


Figura 4.17: Red de titulaciones. Discretización por igual frecuencia y aprendizaje con PC.

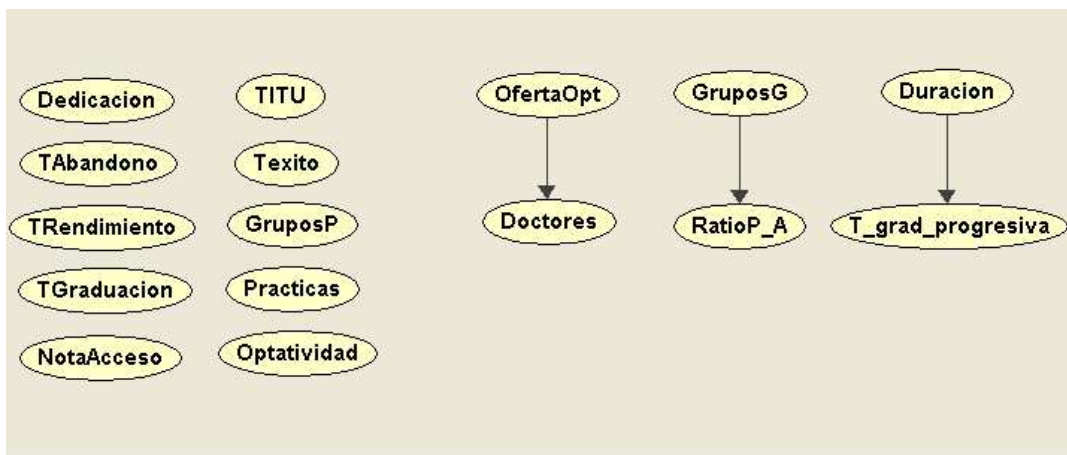


Figura 4.18: Red de titulaciones. Discretización por K-medias y aprendizaje con PC.

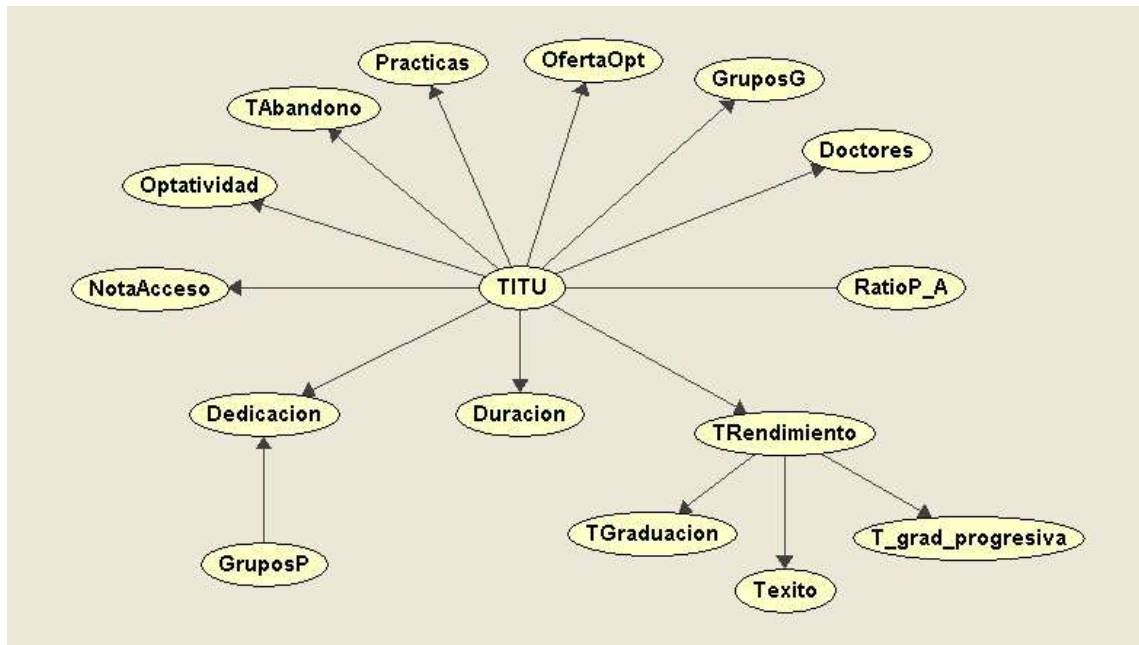


Figura 4.19: Red de titulaciones. Discretización por igual frecuencia y aprendizaje con K2.

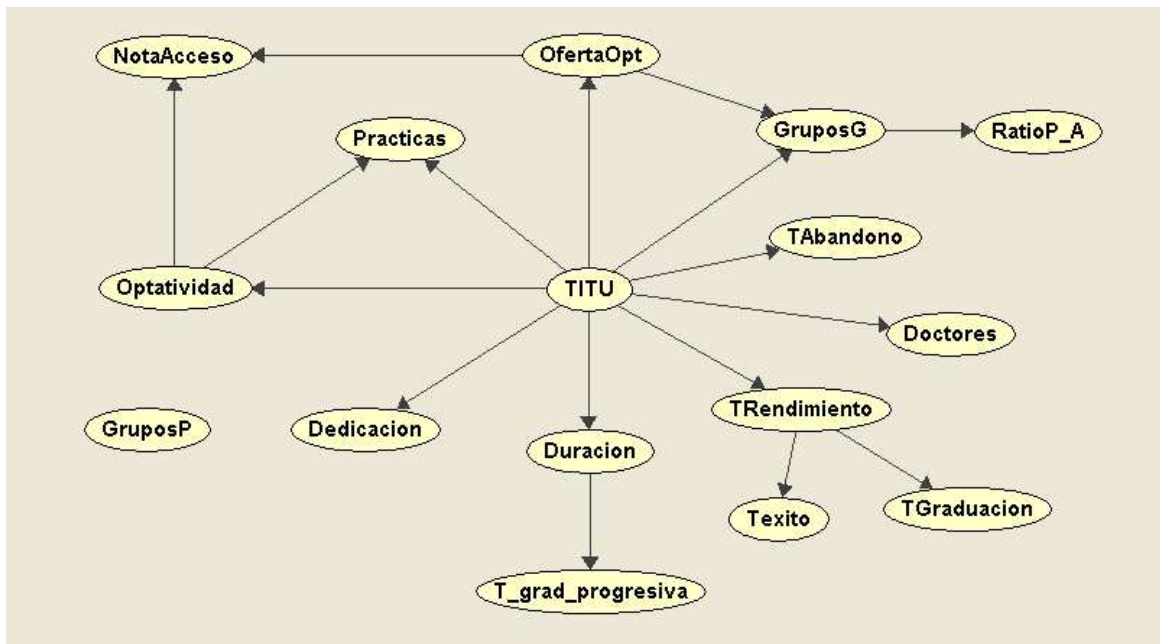


Figura 4.20: Red de titulaciones. Discretización por K-medias y aprendizaje con K2.

DISCRETIZACIÓN	APRENDIZAJE	LOGVERO
Igual frecuencia	PC 95 %	-2.625,7464
K-medias	PC 95 %	-2.444,2100
Igual frecuencia	K2	-1.852,6681
K-medias	K2	-1.799,2598

Tabla 4.24: Logaritmo de la verosimilitud de las redes de titulaciones.

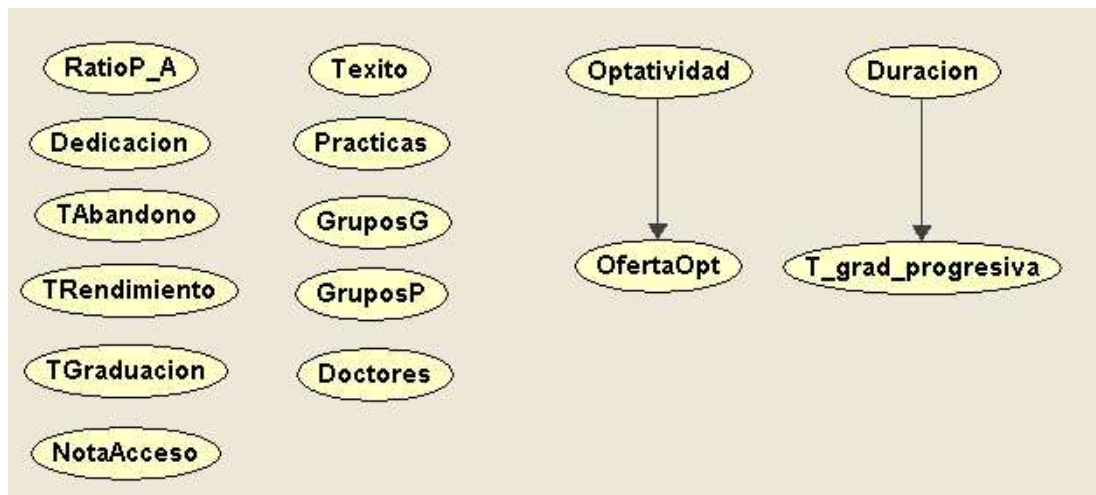


Figura 4.21: Red de titulaciones. Discretización por igual frecuencia y aprendizaje con PC.

Esta red muestra cómo la variable *Titulación* d-separa las variables objetivos del resto de variables que recogen información sobre el perfil del estudio, del profesorado o del alumnado. Esto se debe a que, obviamente, la titulación determina los requerimientos prácticos y optativos del plan de estudios, así como el tamaño de los grupos teóricos. Por lo tanto, optamos por suprimir dicha variable para poder estudiar mejor las dependencias entre el resto. Siguiendo el mismo procedimiento, obtenemos las redes de las figuras 4.21, 4.22, 4.23, 4.24. De la tabla de valores del logaritmo de la verosimilitud, 4.25, obtenemos que la red más adecuada a nuestros datos es la 4.24.

La red 4.24 muestra un gran número de relaciones entre las variables, no pudiéndose encontrar independencias obvias. Las que registran un mayor número de dependencias son el porcentaje de grupos grandes, la oferta de optatividad, el peso práctico en el

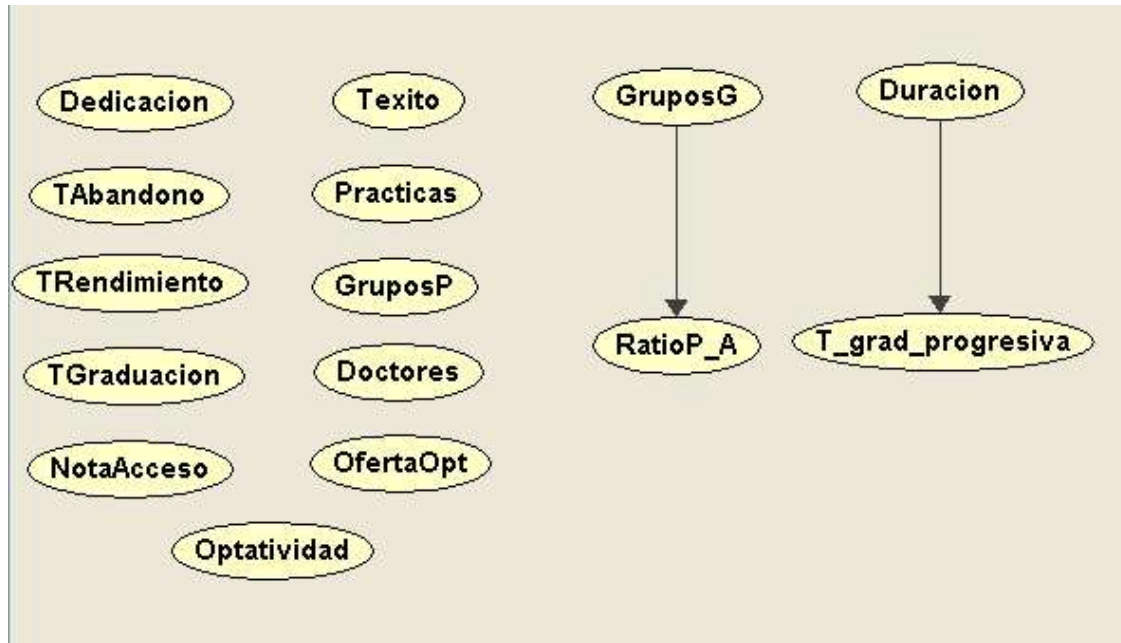


Figura 4.22: Red de titulaciones. Discretización por K-medias y aprendizaje con PC.

DISCRETIZACIÓN	APRENDIZAJE	LOGVERO
Igual frecuencia	PC 95 %	-2.290,1916
K-medias	PC 95 %	-2.138,7737
Igual frecuencia	K2	-1.708,1615
K-medias	K2	-1.647,9312

Tabla 4.25: Logaritmo de la verosimilitud de las redes de titulaciones sin la vble. *Titulación*.

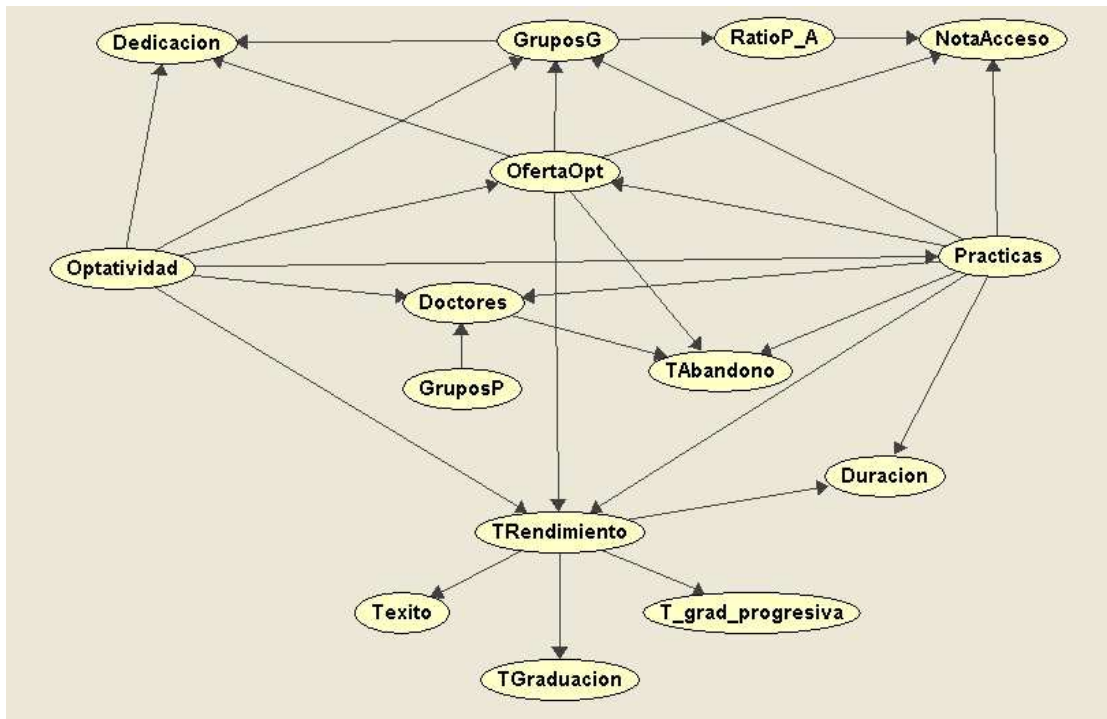


Figura 4.23: Red de titulaciones. Discretización por igual frecuencia y aprendizaje con K2.

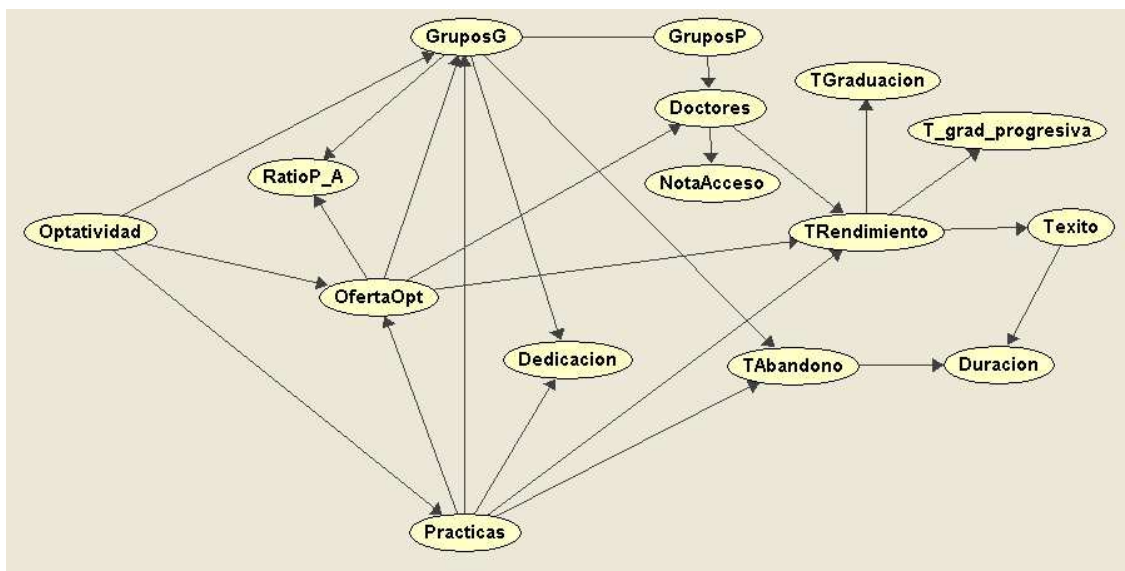


Figura 4.24: Red de titulaciones. Discretización por K-medias y aprendizaje con K2.

Tasa de Abandono	Probabilidad a Priori	Porcentaje de créditos prácticos requeridos				
		[22, 26.5)	[26.5, 34)	[34, 38.5)	[38.5, 46.5)	[46.5, 56]
[0, 9.5)	0.19	0.31	0.21	0.29	0.20	0.15
[9.5, 19.5)	0.27	0.17	0.17	0.19	0.27	0.30
[19.5, 32)	0.26	0.17	0.24	0.16	0.24	0.30
[32, 49)	0.16	0.17	0.17	0.19	0.14	0.16
[49, 66]	0.12	0.17	0.21	0.16	0.14	0.09

Tasa de Abandono	Probabilidad a Priori	Porcentaje de grupos con más de 80 alumnos				
		[0, 4)	[4, 15.5)	[15.5, 40)	[40, 69)	[69, 100]
[0, 9.5)	0.19	0.14	0.37	0.18	0.12	0.19
[9.5, 19.5)	0.27	0.12	0.23	0.31	0.29	0.35
[19.5, 32)	0.26	0.38	0.13	0.26	0.22	0.26
[32, 49)	0.16	0.27	0.13	0.15	0.18	0.09
[49, 66]	0.12	0.09	0.13	0.10	0.19	0.11

Tasa de Abandono	Probabilidad a Priori	Porcentaje de grupos con menos de 20 alumnos				
		[0, 1)	[1, 5.5)	[5.5, 9)	[9, 17)	[17, 24]
[0, 9.5)	0.19	0.18	0.21	0.18	0.20	0.23
[9.5, 19.5)	0.27	0.28	0.29	0.20	0.22	0.28
[19.5, 32)	0.26	0.26	0.24	0.29	0.28	0.22
[32, 49)	0.16	0.1	0.14	0.20	0.19	0.15
[49, 66]	0.12	0.13	0.12	0.12	0.11	0.12

Tabla 4.26: Tasa de abandono en función del peso de las prácticas y el tamaño de los grupos.

plan de estudios y la tasa de rendimiento.

La tasa de abandono de una titulación depende directamente del porcentaje de grupos grandes (con más de 80 alumnos por grupo de teoría) y del peso práctico en el plan de estudios. En la tabla 4.26 se muestran los resultados de la propagación de ambas variables y su efecto sobre la tasa de abandono.

Al estudiar la correlación entre la tasa de abandonos y el peso de las prácticas en el plan de estudios, hay que tener en cuenta que sólo las clases [34, 38.5), [38.5, 46.5) y [46.5, 56] engloban a una variedad representativa de titulaciones, ya que los dos primeros intervalos recogen los resultados únicamente de la licenciatura en derecho y la diplomatura en gestión y administración pública, respectivamente.

De la primera tabla de 4.26 concluimos que para las titulaciones con menos peso práctico las tasas de abandono más probables son las más bajas, mientras que al aumentar el porcentaje de créditos prácticos, esta tasa se desplaza a los intervalos in-

termedios (entre 9.5 % y 32 %). Sin embargo, las probabilidades de abandonos elevados (superiores al 49 %) son más probables en titulaciones con una carga práctica menor, por lo que aumentar la carga práctica en el plan de estudios a un porcentaje superior al 38.5 % de los créditos requeridos mejora la tasa de abandono, puesto que minimiza las probabilidades de tasas altas de abandono.

Observamos en las dos últimas tablas de 4.26, que no hay un patrón claro del comportamiento de la tasa de abandono en función del tamaño de los grupos de la titulación. Los mejores resultados para la tasa de abandono se producen para estudios con un porcentaje de grupos grandes entre el 4 y el 15.5 % del total de grupos teóricos, o bien, para aquellos con más del 17 % de grupos teóricos pequeños. Sin embargo, llama la atención cómo las tasas más bajas de abandono son alcanzadas con más probabilidad en titulaciones con mayor porcentaje de grupos grandes, por lo que deberían realizarse estudios que esclarezcan los motivos de estos resultados.

A continuación realizaremos un estudio de las influencias de las variables escogidas para describir el perfil de la titulación, sobre los resultados de la titulación. En particular nos centraremos en la *tasa de rendimiento*, *coeficiente de duración media*, *tasa de graduación* y *tasa de graduación progresiva*.

La tabla 4.27, muestra los resultados de la propagación del porcentaje de créditos optativos requeridos en el plan de estudios. De ella podemos deducir que la flexibilidad curricular en el plan de estudios no influye ni en el coeficiente de duración media ni en la tasa de graduación; sí lo hace, en cambio, sobre la tasa de rendimiento, disminuyendo las probabilidades de las tasas de rendimiento bajas y provocando un leve incremento en las tasas medias-altas, obteniéndose los mejores resultados al superar los 36 % de créditos optativos sobre el total de créditos requeridos en el plan de estudios. También influye, pero no de modo significativo, en la tasa de graduación progresiva, en la que puede observarse una disminución muy ligera de las tasas pequeñas al superar el 28.5 % de optatividad requerida en el plan de estudios.

La tabla 4.28 recoge los resultados de la propagación del porcentaje de créditos prácticos requeridos en el plan de estudios. Como hemos comentado anteriormente,

Tasa de Rendimiento	Probabilidad a Priori	Optatividad requerida en el plan de estudios				
		[0, 0.215)	[0.215, 0.285)	[0.285, 0.36)	[0.36, 0.48)	[48, 54]
[0.35, 0.42)	0.20	0.21	0.20	0.20	0.19	0.18
[0.42, 0.49)	0.19	0.19	0.18	0.18	0.18	0.19
[0.49, 0.58)	0.20	0.17	0.21	0.22	0.21	0.22
[0.58, 0.70)	0.22	0.22	0.21	0.22	0.23	0.23
[0.70, 0.86]	0.20	0.21	0.19	0.18	0.19	0.18

Coeficiente Duración	Probabilidad a Priori	Optatividad requerida en el plan de estudios				
		[0, 0.215)	[0.215, 0.285)	[0.285, 0.36)	[0.36, 0.48)	[48, 54]
[1, 1.12)	0.13	0.13	0.13	0.13	0.13	0.13
[1.12, 1.38)	0.24	0.24	0.24	0.24	0.24	0.25
[1.38, 1.6)	0.22	0.22	0.22	0.23	0.23	0.23
[1.6, 1.96)	0.21	0.21	0.21	0.21	0.21	0.20
[1.96, 2.73]	0.20	0.20	0.20	0.20	0.20	0.19

Tasa de Graduación	Probabilidad a Priori	Optatividad requerida en el plan de estudios				
		[0, 0.215)	[0.215, 0.285)	[0.285, 0.36)	[0.36, 0.48)	[48, 54]
[0, 0.035)	0.21	0.21	0.21	0.21	0.21	0.21
[0.035, 0.105)	0.27	0.26	0.28	0.28	0.27	0.28
[0.105, 0.235)	0.20	0.20	0.20	0.20	0.21	0.21
[0.235, 0.485)	0.23	0.24	0.23	0.23	0.23	0.23
≥ 0.485	0.08	0.08	0.08	0.08	0.08	0.08

Graduación Progresiva	Probabilidad a Priori	Optatividad requerida en el plan de estudios				
		[0, 0.215)	[0.215, 0.285)	[0.285, 0.36)	[0.36, 0.48)	[48, 54]
[0.61, 0.635)	0.09	0.10	0.09	0.09	0.09	0.09
[0.635, 0.675)	0.16	0.16	0.16	0.15	0.15	0.15
[0.675, 0.72)	0.16	0.15	0.16	0.16	0.15	0.16
[0.72, 0.805)	0.19	0.18	0.19	0.14	0.19	0.20
[0.805, 0.98]	0.41	0.41	0.40	0.40	0.41	0.40

Tabla 4.27: Resultados de la titulación en función del peso de la optatividad en el plan de estudios.

Tasa de Rendimiento	Probabilidad a Priori	Porcentaje de prácticas requeridas				
		[0.22, 0.265]	[0.265, 0.34]	[0.34, 0.385]	[0.385, 0.465]	[0.465, 0.56]
[0.35, 0.42)	0.20	0.19	0.19	0.019	0.19	0.21
[0.42, 0.49)	0.19	0.22	0.19	0.19	0.21	0.17
[0.49, 0.58)	0.20	0.22	0.23	0.20	0.24	0.16
[0.58, 0.70)	0.22	0.19	0.20	0.24	0.19	0.24
[0.70, 0.86]	0.20	0.19	0.19	0.19	0.16	0.22

Coeficiente Duración	Probabilidad a Priori	Porcentaje de prácticas requeridas				
		[0.22, 0.265]	[0.265, 0.34]	[0.34, 0.385]	[0.385, 0.465]	[0.465, 0.56]
[1, 1.12)	0.13	0.13	0.13	0.13	0.13	0.13
[1.12, 1.38)	0.24	0.25	0.23	0.25	0.24	0.24
[1.38, 1.6)	0.22	0.23	0.22	0.23	0.23	0.22
[1.6, 1.96)	0.21	0.20	0.21	0.20	0.21	0.21
[1.96, 2.73]	0.20	0.19	0.21	0.19	0.20	0.20

Tasa de Graduación	Probabilidad a Priori	Porcentaje de prácticas requeridas				
		[0.22, 0.265]	[0.265, 0.34]	[0.34, 0.385]	[0.385, 0.465]	[0.465, 0.56]
[0, 0.035)	0.21	0.22	0.21	0.21	0.22	0.20
[0.035, 0.105)	0.27	0.29	0.28	0.27	0.30	0.25
[0.105, 0.235)	0.20	0.20	0.20	0.21	0.20	0.21
[0.235, 0.485)	0.23	0.22	0.22	0.23	0.21	0.25
≥ 0.485	0.08	0.08	0.08	0.08	0.08	0.09

Graduación Progresiva	Probabilidad a Priori	Porcentaje de prácticas requeridas				
		[0.22, 0.265]	[0.265, 0.34]	[0.34, 0.385]	[0.385, 0.465]	[0.465, 0.56]
[0.61, 0.635)	0.09	0.10	0.09	0.09	0.10	0.09
[0.635, 0.675)	0.16	0.16	0.15	0.15	0.16	0.15
[0.675, 0.72)	0.16	0.17	0.16	0.15	0.17	0.15
[0.72, 0.805)	0.19	0.19	0.19	0.19	0.20	0.18
[0.805, 0.98]	0.41	0.39	0.40	0.41	0.38	0.42

Tabla 4.28: Resultados de la titulación en función del peso práctico en el plan de estudios.

sólo las tres últimas clases representan a una variedad significativa de titulaciones, ya que las dos primeras recogen, exclusivamente, a la licenciatura en derecho y a la diplomatura en gestión y administración pública respectivamente. Centrándonos en el estudio de los valores de la probabilidad de las variables *tasa de rendimiento*, *tasa de graduación* y *tasa de graduación progresiva*, no encontramos una tendencia clara de dichos valores al propagar la variable *Prácticas*; tan sólo destacar que los mejores resultados se observan para planes con una carga práctica superior al 46.5%. El coeficiente de duración media no denota cambios significativos ante los valores de *Prácticas*, aunque sufre un incremento muy pequeño en las probabilidades de coeficientes más altos, debido, quizás, a los proyectos fin de carrera. En todo caso, este inconveniente es mínimo ante los efectos beneficiosos que, tanto en las tasas de graduación y rendimiento, como en la tasa de abandono, conlleva un porcentaje de créditos prácticos en el plan de estudios igual o superior al 46.5% de los créditos totales requeridos.

En tercer lugar estudiamos el efecto que la oferta de optativas puede tener sobre el rendimiento en una titulación. La tabla 4.29 muestra la propagación de la variable *OfertaOpt*. Esta variable está muy determinada por la titulación, ya que, por ejemplo, el intervalo [22.6, 38.16] recoge únicamente los resultados de las distintas especialidades de magisterio ya que las optativas se cofinancian entre los cinco estudios, por lo que tienen una mayor oferta de optativas; una situación análoga se da en la clase [8.77, 22.6), que recoge los resultados de las cuatro especialidades de la ingeniería técnica agrícola y tres de magisterio; la tercera clase representa las especialidades de magisterio, de ingeniería técnica agrícola y de ingeniería técnica informática. Por lo tanto, no podemos discernir si los cambios en las diferentes tasas proceden de la variación en los valores de la tabla 4.29 o por las titulaciones a las que pertenecen.

A continuación propagamos la variable que recoge el porcentaje de grupos con más de 80 alumnos impartidos en la titulación en la cohorte a estudio, obteniendo los resultados de la tabla 4.30. Esta propagación apenas modifica las probabilidades a priori de la tasa de graduación. Tampoco es significativo el efecto que la propagación de *GruposG* tiene sobre las tasas de graduación progresiva y la tasa de rendimiento; en la primera puede observarse una leve disminución de las tasas superiores al 81%

Tasa de Rendimiento	Probabilidad a Priori	Optatividad Ofertada				
		[1, 2.75]	[2.75, 4.94]	[4.94, 8.77]	[8.77, 22.6]	[22.6, 38.16]
[0.35, 0.42)	0.20	0.15	0.23	0.22	0.24	0.16
[0.42, 0.49)	0.19	0.18	0.17	0.20	0.23	0.16
[0.49, 0.58)	0.20	0.27	0.19	0.20	0.15	0.16
[0.58, 0.70)	0.22	0.25	0.21	0.19	0.19	0.28
[0.70, 0.86]	0.20	0.16	0.20	0.19	0.20	0.25

Coeficiente Duración	Probabilidad a Priori	Optatividad Ofertada				
		[1, 2.75]	[2.75, 4.94]	[4.94, 8.77]	[8.77, 22.6]	[22.6, 38.16]
[1, 1.12)	0.13	0.13	0.13	0.13	0.12	0.13
[1.12, 1.38)	0.24	0.25	0.24	0.24	0.23	0.25
[1.38, 1.6)	0.22	0.23	0.22	0.22	0.22	0.22
[1.6, 1.96)	0.21	0.21	0.21	0.21	0.21	0.21
[1.96, 2.73]	0.20	0.19	0.20	0.20	0.21	0.19

Tasa de Graduación	Probabilidad a Priori	Optatividad Ofertada				
		[1, 2.75]	[2.75, 4.94]	[4.94, 8.77]	[8.77, 22.6]	[22.6, 38.16]
[0, 0.035)	0.21	0.20	0.22	0.22	0.23	0.18
[0.035, 0.105)	0.27	0.29	0.27	0.28	0.27	0.23
[0.105, 0.235)	0.20	0.22	0.20	0.19	0.19	0.22
[0.235, 0.485)	0.23	0.22	0.23	0.22	0.22	0.28
≥ 0.485	0.08	0.07	0.08	0.08	0.08	0.09

Graduación Progresiva	Probabilidad a Priori	Optatividad Ofertada				
		[1, 2.75]	[2.75, 4.94]	[4.94, 8.77]	[8.77, 22.6]	[22.6, 38.16]
[0.61, 0.635)	0.09	0.09	0.10	0.10	0.10	0.08
[0.635, 0.675)	0.16	0.14	0.16	0.16	0.18	0.14
[0.675, 0.72)	0.16	0.16	0.16	0.16	0.16	0.13
[0.72, 0.805)	0.19	0.21	0.19	0.19	0.18	0.18
[0.805, 0.98]	0.41	0.41	0.40	0.38	0.38	0.47

Tabla 4.29: Resultados de la titulación en función de la oferta de optativas.

Tasa de Rendimiento	Probabilidad a Priori	Porcentaje grupos grandes				
		[0, 0.04)	[0.04, 0.155)	[0.155, 0.40)	[0.40, 0.69)	[0.69, 1]
[0.35, 0.42)	0.20	0.21	0.19	0.18	0.20	0.21
[0.42, 0.49)	0.19	0.19	0.19	0.18	0.19	0.18
[0.49, 0.58)	0.20	0.19	0.21	0.20	0.20	0.19
[0.58, 0.70)	0.22	0.21	0.22	0.24	0.22	0.21
[0.70, 0.86]	0.20	0.19	0.19	0.20	0.19	0.20

Coeficiente Duración	Probabilidad a Priori	Porcentaje grupos grandes				
		[0, 0.04)	[0.04, 0.155)	[0.155, 0.40)	[0.40, 0.69)	[0.69, 1]
[1, 1.12)	0.13	0.13	0.12	0.13	0.13	0.12
[1.12, 1.38)	0.24	0.20	0.28	0.25	0.22	0.25
[1.38, 1.6)	0.22	0.22	0.23	0.23	0.22	0.22
[1.6, 1.96)	0.21	0.21	0.19	0.21	0.22	0.21
[1.96, 2.73]	0.20	0.24	0.17	0.19	0.19	0.19

Tasa de Graduación	Probabilidad a Priori	Porcentaje grupos grandes				
		[0, 0.04)	[0.04, 0.155)	[0.155, 0.40)	[0.40, 0.69)	[0.69, 1]
[0, 0.035)	0.21	0.22	0.21	0.20	0.21	0.21
[0.035, 0.105)	0.27	0.27	0.27	0.27	0.27	0.27
[0.105, 0.235)	0.20	0.20	0.20	0.21	0.20	0.20
[0.235, 0.485)	0.23	0.23	0.23	0.24	0.23	0.23
≥ 0.485	0.08	0.08	0.08	0.08	0.08	0.08

Graduación Progresiva	Probabilidad a Priori	Porcentaje grupos grandes				
		[0, 0.04)	[0.04, 0.155)	[0.155, 0.40)	[0.40, 0.69)	[0.69, 1]
[0.61, 0.635)	0.09	0.10	0.09	0.09	0.09	0.10
[0.635, 0.675)	0.16	0.16	0.15	0.15	0.16	0.16
[0.675, 0.72)	0.16	0.16	0.16	0.15	0.16	0.16
[0.72, 0.805)	0.19	0.19	0.19	0.19	0.19	0.19
[0.805, 0.98]	0.41	0.40	0.41	0.42	0.40	0.40

Tabla 4.30: Resultados de la titulación en función del porcentaje de grupos grandes.

Tasa de Rendimiento	Probabilidad a Priori	Porcentaje de grupos con menos de 20 alumnos				
		[0, 0.01)	[0.01, 0.055)	[0.055, 0.09)	[0.09, 0.17)	[0.17, 0.24]
[0.35, 0.42)	0.20	0.21	0.19	0.20	0.20	0.18
[0.42, 0.49)	0.19	0.19	0.18	0.20	0.19	0.18
[0.49, 0.58)	0.20	0.19	0.21	0.20	0.20	0.20
[0.58, 0.70)	0.22	0.21	0.23	0.21	0.21	0.25
[0.70, 0.86]	0.20	0.20	0.20	0.19	0.19	0.19

Duración	Probabilidad a Priori	Porcentaje de grupos con menos de 20 alumnos				
		[0, 0.01)	[0.01, 0.055)	[0.055, 0.09)	[0.09, 0.17)	[0.17, 0.24]
[1, 1.12)	0.13	0.13	0.13	0.13	0.13	0.13
[1.12, 1.38)	0.24	0.24	0.25	0.23	0.23	0.25
[1.38, 1.6)	0.22	0.22	0.23	0.22	0.23	0.23
[1.6, 1.96)	0.21	0.21	0.21	0.21	0.21	0.21
[1.96, 2.73]	0.20	0.20	0.19	0.21	0.21	0.19

Tasa de Graduación	Probabilidad a Priori	Porcentaje de grupos con menos de 20 alumnos				
		[0, 0.01)	[0.01, 0.055)	[0.055, 0.09)	[0.09, 0.17)	[0.17, 0.24]
[0, 0.035)	0.21	0.21	0.21	0.22	0.21	0.20
[0.035, 0.105)	0.27	0.27	0.27	0.28	0.28	0.27
[0.105, 0.235)	0.20	0.20	0.21	0.20	0.20	0.21
[0.235, 0.485)	0.23	0.23	0.24	0.23	0.23	0.24
≥ 0.485	0.08	0.08	0.08	0.08	0.08	0.08

Graduación Progresiva	Probabilidad a Priori	Porcentaje de grupos con menos de 20 alumnos				
		[0, 0.01)	[0.01, 0.055)	[0.055, 0.09)	[0.09, 0.17)	[0.17, 0.24]
[0.61, 0.635)	0.09	0.10	0.09	0.10	0.10	0.09
[0.635, 0.675)	0.16	0.16	0.15	0.16	0.16	0.15
[0.675, 0.72)	0.16	0.16	0.15	0.16	0.16	0.15
[0.72, 0.805)	0.19	0.19	0.19	0.19	0.19	0.19
[0.805, 0.98]	0.41	0.40	0.41	0.40	0.40	0.42

Tabla 4.31: Resultados de la titulación en función del porcentaje de grupos pequeños.

cuando los grupos con más de 80 alumnos superan el 40 % de los grupos totales; en la tasa de rendimiento, los mejores resultados se obtienen en el intervalo de 15.5 % al 40 % del total de grupos impartidos. Sí se aprecia una variación en las probabilidades del coeficiente de duración media: mientras las titulaciones con porcentajes de grupos grandes superiores o iguales al 15.5 % tienen un comportamiento parecido, el coeficiente se mejora cuando el porcentaje de grupos grandes se mueve entre el 4 % y el 15.5 %, obteniendo los peores resultados en aquellos estudios con porcentajes inferiores al 4 %.

El porcentaje de grupos con menos de 20 personas influye en la tasa de rendimiento y en la tasa de graduación progresiva; como muestra la tabla 4.31, las probabilidades de tasas de rendimiento bajas disminuyen al aumentar el porcentaje de grupos pequeños

por encima del 9% de los grupos totales, mientras que las tasas más altas son más probables en estudios donde el porcentaje de grupos con menos de 20 personas supera el 17%. También en este caso se nota un ligero ascenso de la probabilidad de obtener tasas de graduación progresiva superiores a 0.81.

La tasa de graduación así como el coeficiente de duración media, no muestran una dependencia destacable con el porcentaje de grupos de teoría pequeños en la titulación.

La propagación del ratio profesor-alumno, no modifica significativamente las probabilidades a priori de las variables a estudio. Los resultados de dicha propagación pueden consultarse en la tabla 4.32, aunque, como ocurría con la oferta de optativas, estos resultados pueden no ser concluyentes, puesto que los ratios vienen dados por las titulaciones; así, en el primer intervalo sólo se recogen los resultados de las ingenierías técnicas agrícolas e informáticas y de las licenciatura en matemáticas, ambientales y psicopedagogía, mientras que las dos últimas clases engloban a las especialidades de magisterio, empresariales, turismo y las licenciaturas de derecho, psicología y filología hispánica.

Una situación análoga se presenta con la variable *porcentaje de créditos impartidos por doctores*, ya que las clases bajas representan a los alumnos de enfermería, magisterio, empresariales, gestión y administración pública y las ingenierías técnicas en informática, mientras que en las clases superiores se recogen los resultados de las licenciaturas en química, matemáticas, ambientales, humanidades o ingeniero agrónomo. Por lo tanto, no pueden sacarse conclusiones de los resultados mostrados en la tabla 4.33, puesto que la variación entre los distintos resultados se deberá a la titulación del alumno en vez de al porcentaje de créditos impartidos por profesores doctores.

Para finalizar, estudiaremos el efecto que el nivel inicial de los alumnos, medido a través de la nota media de acceso a la titulación, tiene sobre los resultados de la citada titulación. Los resultados de dicha propagación se encuentran en la tabla 4.34, que muestra cómo la nota de acceso de los alumnos no tiene una influencia significativa sobre el coeficiente de duración media ni sobre las tasas de graduación y de graduación progresiva. Sin embargo, la tasa de rendimiento de la titulación sí presenta un

Tasa de Rendimiento	Probabilidad a Priori	Ratio profesor-alumno				
		[14.9, 34.55)	[34.55, 50.55)	[50.55, 63.95)	[63.95, 83.7)	[83.7, 115.7]
[0.35, 0.42)	0.20	0.21	0.20	0.19	0.20	0.19
[0.42, 0.49)	0.19	0.20	0.18	0.18	0.19	0.19
[0.49, 0.58)	0.20	0.19	0.20	0.20	0.19	0.20
[0.58, 0.70)	0.22	0.21	0.22	0.23	0.22	0.22
[0.70, 0.86]	0.20	0.19	0.19	0.20	0.20	0.19

Coeficiente Duración	Probabilidad a Priori	Ratio profesor-alumno				
		[14.9, 34.55)	[34.55, 50.55)	[50.55, 63.95)	[63.95, 83.7)	[83.7, 115.7]
[1, 1.12)	0.13	0.13	0.13	0.13	0.13	0.13
[1.12, 1.38)	0.24	0.23	0.24	0.24	0.24	0.24
[1.38, 1.6)	0.22	0.22	0.23	0.22	0.22	0.23
[1.6, 1.96)	0.21	0.21	0.21	0.21	0.21	0.21
[1.96, 2.73]	0.20	0.20	0.19	0.20	0.21	0.20

Tasa de Graduación	Probabilidad a Priori	Ratio profesor-alumno				
		[14.9, 34.55)	[34.55, 50.55)	[50.55, 63.95)	[63.95, 83.7)	[83.7, 115.7]
[0, 0.035)	0.21	0.22	0.21	0.21	0.21	0.21
[0.035, 0.105)	0.27	0.27	0.27	0.27	0.27	0.27
[0.105, 0.235)	0.20	0.20	0.20	0.21	0.20	0.20
[0.235, 0.485)	0.23	0.23	0.23	0.24	0.24	0.23
≥ 0.485	0.08	0.08	0.08	0.08	0.08	0.08

Graduación Progresiva	Probabilidad a Priori	Ratio profesor-alumno				
		[14.9, 34.55)	[34.55, 50.55)	[50.55, 63.95)	[63.95, 83.7)	[83.7, 115.7]
[0.61, 0.635)	0.09	0.10	0.09	0.09	0.09	0.09
[0.635, 0.675)	0.16	0.16	0.15	0.15	0.16	0.15
[0.675, 0.72)	0.16	0.16	0.16	0.15	0.15	0.15
[0.72, 0.805)	0.19	0.19	0.19	0.19	0.19	0.19
[0.805, 0.98]	0.41	0.40	0.41	0.41	0.41	0.40

Tabla 4.32: Resultados en función del ratio profesor-alumno.

Tasa de Rendimiento	Probabilidad a Priori	Porcentaje de créditos impartidos por doctores				
		[0.2, 0.37)	[0.37, 0.555)	[0.555, 0.715)	[0.715, 0.835)	[0.835, 0.93]
[0.35, 0.42)	0.20	0.24	0.17	0.17	0.24	0.18
[0.42, 0.49)	0.19	0.15	0.17	0.21	0.22	0.18
[0.49, 0.58)	0.20	0.16	0.19	0.19	0.20	0.26
[0.58, 0.70)	0.22	0.17	0.26	0.28	0.19	0.20
[0.70, 0.86]	0.20	0.28	0.22	0.15	0.15	0.18

Coeficiente Duración	Probabilidad a Priori	Porcentaje de créditos impartidos por doctores				
		[0.2, 0.37)	[0.37, 0.555)	[0.555, 0.715)	[0.715, 0.835)	[0.835, 0.93]
[1, 1.12)	0.13	0.13	0.13	0.13	0.12	0.13
[1.12, 1.38)	0.24	0.24	0.24	0.24	0.23	0.24
[1.38, 1.6)	0.22	0.22	0.23	0.23	0.22	0.23
[1.6, 1.96)	0.21	0.21	0.21	0.21	0.21	0.21
[1.96, 2.73]	0.20	0.19	0.20	0.21	0.20	0.20

Tasa de Graduación	Probabilidad a Priori	Porcentaje de créditos impartidos por doctores				
		[0.2, 0.37)	[0.37, 0.555)	[0.555, 0.715)	[0.715, 0.835)	[0.835, 0.93]
[0, 0.035)	0.21	0.21	0.19	0.21	0.23	0.21
[0.035, 0.105)	0.27	0.25	0.25	0.27	0.29	0.29
[0.105, 0.235)	0.20	0.17	0.21	0.23	0.20	0.20
[0.235, 0.485)	0.23	0.26	0.26	0.23	0.20	0.22
≥ 0.485	0.08	0.10	0.08	0.07	0.08	0.08

Graduación Progresiva	Probabilidad a Priori	Porcentaje de créditos impartidos por doctores				
		[0.2, 0.37)	[0.37, 0.555)	[0.555, 0.715)	[0.715, 0.835)	[0.835, 0.93]
[0.61, 0.635)	0.09	0.10	0.09	0.09	0.10	0.09
[0.635, 0.675)	0.16	0.16	0.14	0.15	0.17	0.15
[0.675, 0.72)	0.16	0.15	0.14	0.15	0.17	0.16
[0.72, 0.805)	0.19	0.17	0.19	0.20	0.19	0.20
[0.805, 0.98]	0.41	0.42	0.44	0.41	0.36	0.40

Tabla 4.33: Resultados de la titulación en función del perfil docente.

Tasa de Rendimiento	Probabilidad a Priori	Nota media de acceso				
		[5.69, 5.92)	[5.92, 6.12)	[6.12, 6.43)	[6.43, 6.95)	[6.95, 7.39]
[0.35, 0.42)	0.20	0.21	0.20	0.19	0.20	0.22
[0.42, 0.49)	0.19	0.17	0.19	0.19	0.20	0.17
[0.49, 0.58)	0.20	0.18	0.20	0.21	0.20	0.18
[0.58, 0.70)	0.22	0.21	0.23	0.23	0.23	0.19
[0.70, 0.86]	0.20	0.23	0.19	0.18	0.17	0.23

Duración	Coeficiente a Priori	Nota media de acceso				
		[5.69, 5.92)	[5.92, 6.12)	[6.12, 6.43)	[6.43, 6.95)	[6.95, 7.39]
[1, 1.12)	0.13	0.13	0.13	0.13	0.13	0.13
[1.12, 1.38)	0.24	0.24	0.24	0.24	0.24	0.24
[1.38, 1.6)	0.22	0.22	0.23	0.23	0.23	0.22
[1.6, 1.96)	0.21	0.21	0.21	0.21	0.21	0.21
[1.96, 2.73]	0.20	0.20	0.20	0.20	0.20	0.20

Tasa de Graduación	Probabilidad a Priori	Nota media de acceso				
		[5.69, 5.92)	[5.92, 6.12)	[6.12, 6.43)	[6.43, 6.95)	[6.95, 7.39]
[0, 0.035)	0.21	0.21	0.21	0.21	0.21	0.21
[0.035, 0.105)	0.27	0.26	0.27	0.28	0.28	0.26
[0.105, 0.235)	0.20	0.20	0.21	0.21	0.21	0.19
[0.235, 0.485)	0.23	0.25	0.23	0.22	0.22	0.25
≥ 0.485	0.08	0.09	0.08	0.08	0.08	0.09

Graduación Progresiva	Probabilidad a Priori	Nota media de acceso				
		[5.69, 5.92)	[5.92, 6.12)	[6.12, 6.43)	[6.43, 6.95)	[6.95, 7.39]
[0.61, 0.635)	0.09	0.09	0.09	0.09	0.09	0.10
[0.635, 0.675)	0.16	0.15	0.16	0.16	0.16	0.16
[0.675, 0.72)	0.16	0.15	0.16	0.16	0.16	0.15
[0.72, 0.805)	0.19	0.18	0.19	0.19	0.19	0.18
[0.805, 0.98]	0.41	0.42	0.40	0.40	0.40	0.41

Tabla 4.34: Resultados de la titulación en función de la nota media de acceso a la titulación.

ligero aumento en las probabilidades de los valores superiores al 58 % de los créditos matriculados al subir la nota media de acceso por encima de 5.92 puntos.

Por lo tanto, del análisis de los resultados de las titulaciones en función de las variables informativas escogidas, se concluye que:

- Los mejores resultados en la tasa de rendimiento se consiguen aumentando el porcentaje de optatividad en el plan de estudios por encima del 36 % de la carga total del plan de estudios.
- Para mejorar la tasa de graduación y de rendimiento a través del peso práctico en el plan de estudios, los créditos prácticos deberían superar el 46.5 % de los

créditos requeridos. Con esta medida además se optimiza la tasa de abandono.

- Un porcentaje de grupos con más de 80 alumnos situado entre el 4% y el 40% mejora los resultados de la tasa de rendimiento y el coeficiente de duración media de los estudios, reduciéndose la tasa de abandono.
- Un aumento del porcentaje de grupos con menos de 20 alumnos por encima del 17% optimiza los resultados de la tasa de abandono, de rendimiento y de graduación progresiva.

Los efectos sobre los resultados de las titulaciones de estas medidas actuando conjuntamente pueden observarse en la tabla 4.35. La tabla muestra cómo aplicando cada una de las medidas propuestas va reduciéndose la probabilidad de tasas de rendimiento bajas del 0.39 inicial a 0.31 y aumentándose la probabilidad de las tasas de rendimiento altas de 0.42 a 0.53. Menos destacadas son las consecuencias sobre la tasa de graduación, que mejora la probabilidad de las tasas mayores del 31% al 36%, disminuyendo la probabilidad de tasas de graduación bajas. El efecto que estas medidas tienen sobre el coeficiente de duración media es muy pequeño, pero reduce levemente la probabilidad de obtener coeficientes superiores a 1.96. En cuanto a los abandonos, no se consigue aumentar la probabilidad de obtener tasas bajas, pero se disminuye la de las tasas más altas del 28% al 23%.

La manipulación del porcentaje de optatividad no introduce ninguna mejora en los resultados obtenidos, para ninguno de los intervalos de optatividad propagado.

Tasa de Rendimiento	Probabilidad a Priori	Grupos > 80 [15.5 %, 40 %)	Grupos < 20 ≥ 17 %	Peso Práctico ≥ 46.5 %
[0.35, 0.42)	0.20	0.19	0.17	0.16
[0.42, 0.49)	0.19	0.18	0.18	0.15
[0.49, 0.58)	0.20	0.20	0.20	0.15
[0.58, 0.70)	0.22	0.24	0.26	0.30
[0.70, 0.86]	0.20	0.20	0.19	0.23

Coficiente Duración	Probabilidad a Priori	Grupos > 80 [15.5 %, 40 %)	Grupos < 20 ≥ 17 %	Peso Práctico ≥ 46.5 %
[1, 1.12)	0.13	0.13	0.13	0.13
[1.12, 1.38)	0.24	0.25	0.25	0.25
[1.38, 1.6)	0.22	0.23	0.23	0.22
[1.6, 1.96)	0.21	0.21	0.21	0.21
[1.96, 2.73]	0.20	0.19	0.19	0.19

Tasa de Graduación	Probabilidad a Priori	Grupos > 80 [15.5 %, 40 %)	Grupos < 20 ≥ 17 %	Peso Práctico ≥ 46.5 %
[0, 0.035)	0.21	0.20	0.20	0.18
[0.035, 0.105)	0.27	0.27	0.26	0.23
[0.105, 0.235)	0.20	0.21	0.22	0.23
[0.235, 0.485)	0.23	0.24	0.24	0.28
≥ 0.485	0.08	0.08	0.08	0.08

Tasa de Abandono	Probabilidad a Priori	Grupos > 80 [15.5 %, 40 %)	Grupos < 20 ≥ 17 %	Peso Práctico ≥ 46.5 %
[0, 0.095)	0.19	0.18	0.18	0.15
[0.095, 0.195)	0.27	0.31	0.31	0.31
[0.195, 0.32)	0.26	0.26	0.26	0.31
[0.32, 0.49)	0.16	0.15	0.15	0.15
[0.49, 0.66]	0.12	0.10	0.10	0.08

Tabla 4.35: Efectos de las medidas recomendadas.

4.3. Extracción de perfiles.

En este apartado intentaremos dar un perfil de las asignaturas y titulaciones con mejores resultados.

4.3.1. Perfil de asignaturas con mayor rendimiento.

Vamos a extraer los tres perfiles de las asignaturas con mejores tasas de rendimiento diferenciando por el carácter de la asignatura.

Para las asignaturas troncales y obligatorias, la tabla 4.36 recoge los tres perfiles más probables de las asignaturas con rendimiento más alto. Hay que notar que los perfiles de ambas tablas son muy similares; en los seis perfiles la docencia la imparte un único profesor, con los créditos impartidos por docentes doctores permanentes tomando los valores más altos y más del 94 % del alumnado a tiempo completo. Los grupos de teoría tienen 25 alumnos o menos en los perfiles más probables, y entre 45 y 74 alumnos en los siguientes. El tamaño de los grupos de prácticas toma valores en los tres primeros intervalos, siendo los perfiles más probables aquellos con menos de 23 alumnos por grupo práctico.

Como hemos estudiado, no todas estas variables tienen efecto sobre la tasa de rendimiento de las asignaturas troncales y obligatorias. Si estudiamos los perfiles sólo con las variables que influyen sobre esta tasa, obtenemos los resultados de la tabla 4.37, que ponen de manifiesto que el perfil de una asignatura con valores altos de rendimiento (superiores al 69.5 % de los matriculados) es el de asignaturas con grupos pequeños (menos de 25 alumnos o menos por grupo de teoría y 23 alumnos como máximo en los grupos prácticos) y nota de acceso media-alta.

La situación cambia en las asignaturas optativas, en las que los perfiles de las asignaturas con tasa de rendimiento superior al 92 % son casi opuestos a los perfiles de las asignaturas con tasas entre el 82 % y el 92 % de los matriculados (ver tabla 4.38). Los perfiles de las asignaturas con mayor tasa de rendimiento corresponden a asignaturas de segundo ciclo, con grupos teóricos y prácticos muy pequeños (menos de

VARIABLES	Tasa de rendimiento superior al 84.5 %		
Titulación	Materiales	Enfermería	Enfermería
Curso	4	1	1
Alum_GT	< 25.5	[44.75, 74.5)	[44.75, 74.5)
Alum_GP	< 23.9	[41.65, 68.5)	[23.9, 41.65)
NºProfesores	1	1	1
EvaluaProf	< 1.33	[3.01, 4.245)	[3.01, 4.245)
Creditos_Doctor	≥ 87 %	≥ 87 %	≥ 87 %
ContratoProf	≥ 89.5 %	≥ 89.5 %	≥ 89.5 %
MediaAcceso	[5.33, 5.915)	≥ 6.88	≥ 6.88
P80acceso	[5.46, 6.30)	≥ 7.61	≥ 7.61
TC	≥ 94.5 %	≥ 94.5 %	≥ 94.5 %
Probabilidad	0.000212	0.000182	0.000166

VARIABLES	Tasa de rendimiento entre 69.5 % y 84.5 %		
Titulación	Materiales	Enfermería	Enfermería
Curso	4	1	1
Alum_GT	< 25.5	[44.75, 74.5)	[44.75, 74.5)
Alum_GP	< 23.9	[41.65, 68.5)	[23.9, 41.65)
NºProfesores	1	1	1
EvaluaProf	< 1.33	[3.01, 4.245)	[3.01, 4.245)
Creditos_Doctor	≥ 87 %	≥ 87 %	≥ 87 %
ContratoProf	≥ 89.5 %	≥ 89.5 %	≥ 89.5 %
MediaAcceso	[5.33, 5.915)	≥ 6.88	≥ 6.88
P80acceso	[5.46, 6.30)	≥ 7.61	≥ 7.61
TC	< 55 %	≥ 94.5 %	≥ 94.5 %
Probabilidad	0.000132	0.000124	0.000113

Tabla 4.36: Perfiles asignaturas obligatorias con mayor rendimiento.

VARIABLES	Tasa de rendimiento superior al 84.5 %		
Alum_GT	< 25.5	< 25.5	< 25.5
Alum_GP	< 23.9	< 23.9	< 23.9
MediaAcceso	[5.92, 6.17)	[6.47, 6.88)	[6.17, 6.47)
P80acceso	[6.30, 6.65)	[7.3, 7.61)	[6.65, 7.73)
Probabilidad	0.037730	0.0033802	0.027684

VARIABLES	Tasa de rendimiento entre 69.5 % y 84.5 %		
Alum_GT	< 25.5	< 25.5	< 25.5
Alum_GP	< 23.9	< 23.9	< 23.9
MediaAcceso	[5.92, 6.17)	[6.47, 6.88)	[6.17, 6.47)
P80acceso	[6.30, 6.65)	[7.3, 7.61)	[6.65, 7.73)
Probabilidad	0.035817	0.034134	0.0027645

Tabla 4.37: Perfiles de las asignaturas obligatorias con mayor rendimiento.

Variables	Tasa de rendimiento superior al 92 %		
Ciclo	2	2	1
Alum_GT	< 10	< 10	< 10
Alum_GP	< 9	< 9	< 9
MediaAcceso	[6.69, 8.03)	[6.69, 8.03)	[6.69, 8.03)
P80acceso	[7.22, 8.57)	[7.22, 8.57)	[7.22, 8.57)
TC	100 %	≥ 77 %	100 %
Probabilidad	0.015467	0.007655	0.007423

Variables	Tasa de rendimiento entre 82 % y 92 %		
Ciclo	1	1	1
Alum_GT	≥ 51	≥ 51	≥ 51
Alum_GP	≥ 38	≥ 38	≥ 38
MediaAcceso	[5.32, 6)	[5.32, 6)	[6, 6.2)
P80acceso	[5.37, 6.28)	[5.37, 6.28)	[6.28, 6.61)
TC	[94, 100)	[88, 94)	[88, 94)
Probabilidad	0.015467	0.007655	0.007423

Tabla 4.38: Perfiles de asignaturas optativas con mayor rendimiento.

10 alumnos) y cuyos alumnos tienen nota de acceso alta y están dedicados al estudio a tiempo completo en su mayoría.

4.3.2. Perfil de titulaciones con mayor rendimiento.

El objetivo ahora es obtener un perfil de las titulaciones que obtienen mejores resultados, para ello extraeremos los tres perfiles más probables de aquellos planes con resultados óptimos en la tasa de rendimiento (superior al 70 % de los créditos matriculados), en su tasa de graduación (superior al 48.5 %) y en el tiempo empleado por sus alumnos en graduarse (coeficiente de duración media inferior a 1.115). Para cada una de estas variables, obtenemos los perfiles de las titulaciones en las variables estudiadas en la sección 4.2, excepto las variables *OfertaOpt*, *Doctores* y *Ratio_PA*, que, como vimos en dicho apartado, están determinadas por los planes a estudio. La tabla 4.39 muestra cómo lo único que varía en los perfiles es la nota de acceso, que toma valores en los tres intervalos más bajos. El perfil de las titulaciones con mejores resultados coincide con titulaciones con muy poca optatividad en el plan de estudios y porcentaje alto de créditos prácticos, con grupos teóricos en su mayoría grandes y un porcentaje muy pequeño de grupos teóricos con menos de 20 alumnos.

Variables	Tasa de rendimiento superior al 70 %		
Optatividad	< 21.5 %	< 21.5 %	< 21.5 %
Practicas	≥ 46.5 %	≥ 46.5 %	≥ 46.5 %
Grupos Grandes	≥ 69 %	≥ 69 %	≥ 69 %
Grupos Pequeños	< 1 %	< 1 %	< 1 %
Nota de Acceso	[5.69, 5.92)	[6.12, 6.43)	[5.92, 6.12)
Dedicación	[59.59, 63.79)	[59.59, 63.79)	[59.59, 63.79)
Tasa Abandono	[9.5, 19.5)	[9.5, 19.5)	[9.5, 19.5)
Probabilidad	0.004963	0.003382	0.003306

Variables	Coeficiente de duración media inferior a 1.115		
Optatividad	< 21.5 %	< 21.5 %	< 21.5 %
Practicas	≥ 46.5 %	≥ 46.5 %	≥ 46.5 %
Grupos Grandes	≥ 69 %	≥ 69 %	≥ 69 %
Grupos Pequeños	< 1 %	< 1 %	< 1 %
Nota de Acceso	[6.12, 6.43)	[5.69, 5.92)	[5.92, 6.12)
Dedicación	[59.59, 63.79)	[59.59, 63.79)	[59.59, 63.79)
Tasa Abandono	[9.5, 19.5)	[9.5, 19.5)	[9.5, 19.5)
Probabilidad	0.003317	0.003193	0.002951

Variables	Tasa de graduación superior al 48.5 %		
Optatividad	< 21.5 %	< 21.5 %	< 21.5 %
Practicas	≥ 46.5 %	≥ 46.5 %	≥ 46.5 %
Grupos Grandes	≥ 69 %	≥ 69 %	≥ 69 %
Grupos Pequeños	< 1 %	< 1 %	< 1 %
Nota de Acceso	[6.12, 6.43)	[5.69, 5.92)	[5.92, 6.12)
Dedicación	[59.59, 63.79)	[59.59, 63.79)	[59.59, 63.79)
Tasa Abandono	[9.5, 19.5)	[9.5, 19.5)	[9.5, 19.5)
Probabilidad	0.004025	0.003599	0.003286

Variables	Tasa de graduación progresiva superior al 80.5 %		
Optatividad	< 21.5 %	< 21.5 %	< 21.5 %
Practicas	≥ 46.5 %	≥ 46.5 %	≥ 46.5 %
Grupos Grandes	≥ 69 %	≥ 69 %	≥ 69 %
Grupos Pequeños	< 1 %	< 1 %	< 1 %
Nota de Acceso	[6.12, 6.43)	[5.69, 5.92)	[5.92, 6.12)
Dedicación	[59.59, 63.79)	[59.59, 63.79)	[59.59, 63.79)
Tasa Abandono	[9.5, 19.5)	[9.5, 19.5)	[9.5, 19.5)
Probabilidad	0.003721	0.003328	0.003086

Tabla 4.39: Perfiles de titulaciones con mayor rendimiento.

Tasa de Rendimiento	Probabilidad a Priori	Grupos > 80 ≥ 69 %	Grupos < 20 < 1 %	Peso Práctico ≥ 46.5 %
[0.35, 0.42)	0.20	0.21	0.22	0.24
[0.42, 0.49)	0.19	0.18	0.18	0.16
[0.49, 0.58)	0.20	0.19	0.18	0.15
[0.58, 0.70)	0.22	0.21	0.21	0.23
[0.70, 0.86]	0.20	0.20	0.20	0.22

Coefficiente Duración	Probabilidad a Priori	Grupos > 80 ≥ 69 %	Grupos < 20 < 1 %	Peso Práctico ≥ 46.5 %
[1, 1.12)	0.13	0.12	0.12	0.12
[1.12, 1.38)	0.24	0.25	0.25	0.26
[1.38, 1.6)	0.22	0.22	0.22	0.22
[1.6, 1.96)	0.21	0.21	0.21	0.22
[1.96, 2.73]	0.20	0.19	0.19	0.18

Tasa de Graduación	Probabilidad a Priori	Grupos > 80 ≥ 69 %	Grupos < 20 < 1 %	Peso Práctico ≥ 46.5 %
[0, 0.035)	0.21	0.21	0.21	0.21
[0.035, 0.105)	0.27	0.27	0.27	0.27
[0.105, 0.235)	0.20	0.20	0.20	0.20
[0.235, 0.485)	0.23	0.23	0.23	0.25
≥ 0.485	0.08	0.08	0.08	0.09

Tasa de Abandono	Probabilidad a Priori	Grupos > 80 ≥ 69 %	Grupos < 20 < 1 %	Peso Práctico ≥ 46.5 %
[0, 0.095)	0.19	0.19	0.19	0.13
[0.095, 0.195)	0.27	0.35	0.35	0.48
[0.195, 0.32)	0.26	0.26	0.26	0.29
[0.32, 0.49)	0.16	0.09	0.09	0.03
[0.49, 0.66]	0.12	0.11	0.11	0.06

Tabla 4.40: Efectos de los valores obtenidos en los perfiles.

Este perfil discrepa con los resultados vistos en el apartado 4.2, en el que se vio cómo los resultados de las titulaciones mejoraban al disminuir el tamaño de los grupos. La tabla 4.40 muestra cómo influiría, sobre los resultados de un estudio, la propagación de grupos grandes de teoría superiores al 69% de los grupos impartidos, menos del 1% de grupos pequeños y un peso práctico superior al 46.5% de los créditos del plan de estudios, tal y como muestran los perfiles obtenidos. Los resultados obtenidos son peores que en la tabla 4.35, salvo en la tasa de abandono.

La explicación de los perfiles de la tabla 4.39 reside en las características de las titulaciones con mejores resultados: si propagamos las mayores tasas de rendimiento en la red 4.20, obtenemos que las titulaciones más probables son:

- Enfermería (probabilidad 0.15),
- Magisterio, especialidad en educación infantil (probabilidad 0.06),
- Magisterio, especialidad en educación musical (probabilidad 0.06).

Al propagar las titulaciones con menor coeficiente de duración media, obtenemos:

- Enfermería (probabilidad 0.10),
- Diplomado en Turismo (probabilidad 0.10),
- Licenciado en Filología Inglesa (0.10).

Por último, las titulaciones con mayor tasa de graduación son:

- Enfermería (probabilidad 0.10),
- Magisterio, especialidad en educación infantil (probabilidad 0.05),

Estas titulaciones se caracterizan por tener (salvo en el caso de Filología inglesa) grupos grandes de teoría, lo que provoca los perfiles obtenidos.

Capítulo 5

PREDICCIÓN EN ESTADÍSTICA UNIVERSITARIA

La predicción de magnitudes en estadística universitaria puede ser una herramienta de gran utilidad para la toma de decisiones en el ambiente de la gestión universitaria. Es claro el interés que supondría poder conocer de antemano valores de parámetros como el número de alumnos en una asignatura, el valor de la tasa de rendimiento, de la tasa de éxito o de cualquier otro indicador que pueda afectar a la organización e incluso a la financiación de la universidad.

Un primer acercamiento a la predicción de estas cantidades podría ser el analizar series temporales, pero diversos factores como que la disponibilidad de datos es algo relativamente reciente, así como los continuos cambios en planes de estudio, que afectan a factores determinantes en los indicadores, dificultan la construcción de dichas series.

Una alternativa más viable podría ser construir modelos de regresión con los datos disponibles, radicando aquí la principal dificultad en la heterogeneidad de los datos, pues algunas de las variables son de naturaleza discreta o cualitativa (asignatura, titulación, etc.) mientras que otras son continuas (tasas, etc).

En este capítulo proponemos un modelo de regresión mixto (con variables independientes tanto discretas como continuas) basado en el uso de un tipo especial de

red bayesiana, conocido como *modelo Bayes ingenuo*. La motivación de la elección de tal modelo radica en el buen rendimiento mostrado en problemas de clasificación [18], que pueden entenderse como problemas de regresión donde la variable dependiente es cualitativa.

Los modelos tipo Bayes ingenuo se han empleado para resolver problemas de regresión, pero bajo la suposición de normalidad de la distribución conjunta de las variables dependiente e independientes [20]. Cuando la suposición de normalidad no se cumple, el problema de la regresión usando modelos tipo Bayes ingenuo se ha enfocado usando densidades kernel para representar las distribuciones condicionadas de la correspondiente red bayesiana [17], pero los resultados obtenidos son pobres. Además, el uso de kernels en el contexto de redes bayesianas introduce una alta complejidad en el modelo, lo que puede ser problemático especialmente porque los algoritmos estándar para realizar los cálculos en redes bayesianas [42] no son aplicables a los kernels. Finalmente, una restricción común tanto a los modelos normales como a los basados en kernels es que requieren que todas las variables sean continuas.

Como mencionamos anteriormente, estamos interesados en problemas donde algunas de las variables independientes son discretas o cualitativas mientras que otras son continuas, por lo que la suposición de normalidad no es asumible en ningún caso. Proponemos por tanto un modelo tipo Bayes ingenuo basado en la aproximación de la distribución conjunta de las variables dependiente e independientes por una *Mixtura de Exponenciales Truncadas*, que abreviaremos por MTE. El modelo MTE [29] fue propuesto en el contexto de las redes bayesianas como solución a la presencia de variables discretas y continuas simultáneamente, mostrando buen comportamiento como modelo exacto y también como aproximación de otras distribuciones de probabilidad [3, 4].

5.1. Redes bayesianas para regresión

El antecedente al uso de redes bayesianas para regresión lo encontramos en el problema de la clasificación. Una red bayesiana puede usarse como clasificador si está for-

mada por una variable *clase o dependiente* C y un conjunto de variables *predictoras o independientes* X_1, \dots, X_n , de forma que un individuo con características observadas x_1, \dots, x_n será clasificado como miembro de la clase c^* obtenida como

$$c^* = \arg \max_{c \in \Omega_C} p(c|x_1, \dots, x_n) , \quad (5.1)$$

donde Ω_C denota el soporte de la variable C . De forma similar, una red bayesiana puede emplearse para regresión, es decir, cuando la variable C es continua. Sin embargo, en este caso el objetivo es obtener la distribución a posteriori de la variable dependiente dadas las observaciones x_1, \dots, x_n , y una vez que esta distribución ha sido calculada, se puede dar una predicción numérica de la variable C usando su media, su mediana o su moda.

Obsérvese que $p(c|x_1, \dots, x_n)$ es proporcional a $p(c) \times p(x_1, \dots, x_n|c)$, y por lo tanto resolver el problema de la regresión requiere especificar una distribución n -dimensional para X_1, \dots, X_n dada C . Haciendo uso de la factorización impuesta por la estructura de la red bayesiana, este problema puede simplificarse. El caso extremo se corresponde con la llamada estructura *Bayes ingenuo* [18, 14], donde todas las variables predictoras se consideran condicionalmente independientes dada C . Un ejemplo de esta estructura puede verse en la figura 5.1.

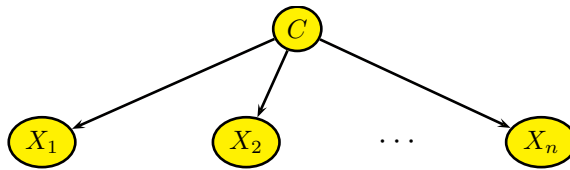


Figura 5.1: Estructura de un modelo de regresión tipo Bayes ingenuo

La fuerte suposición de independencia en la que se basa el modelo Bayes ingenuo se compensa en cierto modo con la reducción en el número de parámetros del modelo, dado que en este caso se cumple que

$$p(c|x_1, \dots, x_n) = p(c) \prod_{i=1}^n p(x_i|c) , \quad (5.2)$$

lo que quiere decir que, en lugar de una distribución condicionada n -dimensional, se requieren n distribuciones condicionadas unidimensionales.

5.2. El modelo MTE

El modelo MTE (del inglés mixture of truncated exponentials) queda definido por su correspondiente potencial y densidad de la siguiente forma [29]:

Definición 17. (Potencial MTE) *Sea \mathbf{X} un vector aleatorio mixto n -dimensional. Sean $\mathbf{Y} = (Y_1, \dots, Y_d)$ y $\mathbf{Z} = (Z_1, \dots, Z_c)$ sus partes discreta y continua respectivamente, con $c + d = n$. Se dice que una función $f : \Omega_{\mathbf{X}} \mapsto \mathbb{R}_0^+$ es un potencial tipo mixtura de exponenciales truncadas (potencial MTE) si verifica una de las siguientes condiciones:*

i. $\mathbf{Y} = \emptyset$ y f puede escribirse como

$$f(\mathbf{x}) = f(\mathbf{z}) = a_0 + \sum_{i=1}^m a_i \exp \left\{ \sum_{j=1}^c b_i^{(j)} z_j \right\} \quad (5.3)$$

para todo $\mathbf{z} \in \Omega_{\mathbf{Z}}$, donde a_i , $i = 0, \dots, m$ y $b_i^{(j)}$, $i = 1, \dots, m$, $j = 1, \dots, c$ son números reales.

ii. $\mathbf{Y} = \emptyset$ y hay una partición D_1, \dots, D_k de $\Omega_{\mathbf{Z}}$ en hipercubos tal que f está definida como

$$f(\mathbf{x}) = f(\mathbf{z}) = f_i(\mathbf{z}) \quad \text{si } \mathbf{z} \in D_i ,$$

donde cada f_i , $i = 1, \dots, k$ se puede escribir como en la ecuación (5.3).

iii. $\mathbf{Y} \neq \emptyset$ y para cada valor fijo $\mathbf{y} \in \Omega_{\mathbf{Y}}$, $f_{\mathbf{y}}(\mathbf{z}) = f(\mathbf{y}, \mathbf{z})$ puede definirse como en ii.

Ejemplo 8. La función ϕ definida como

$$\phi(z_1, z_2) = \begin{cases} 2 + e^{3z_1+z_2} + e^{z_1+z_2} & \text{si } 0 < z_1 \leq 1, 0 < z_2 < 2 \\ 1 + e^{z_1+z_2} & \text{si } 0 < z_1 \leq 1, 2 \leq z_2 < 3 \\ \frac{1}{4} + e^{2z_1+z_2} & \text{si } 1 < z_1 < 2, 0 < z_2 < 2 \\ \frac{1}{2} + 5e^{z_1+2z_2} & \text{si } 1 < z_1 < 2, 2 \leq z_2 < 3 \end{cases}$$

es un potencial MTE, dado que todas sus partes lo son.

Definición 18. (Densidad MTE) *Un potencial MTE f es una densidad MTE si*

$$\sum_{\mathbf{y} \in \Omega_{\mathbf{Y}}} \int_{\Omega_{\mathbf{Z}}} f(\mathbf{y}, \mathbf{z}) d\mathbf{z} = 1 .$$

Una *densidad condicionada MTE* se puede especificar dividiendo el dominio de las variables condicionantes y dando en cada región resultante una densidad MTE para la variable condicionada. Tal es la representación de las densidades condicionadas adoptada en [29], donde también se propone una estructura de datos para manejarla, llamada *árbol de probabilidad mixto*, que puede definirse formalmente como sigue:

Definición 19. (Árbol de probabilidad mixto) *Se dice que un árbol \mathcal{T} es un árbol de probabilidad mixto o árbol mixto, si cumple las siguientes condiciones:*

- i. *Cada nodo interior representa una variable aleatoria (discreta o continua).*
- ii. *Cada arco saliente de una variable Z se etiqueta con un intervalo de valores de Z , de forma que el dominio de Z es la unión de los intervalos correspondientes a arcos que salen de Z .*
- iii. *Cada variable discreta o cualitativa tiene un número de arcos salientes igual a la cantidad de posibles valores que puede tomar.*
- iv. *Cada nodo hoja contiene un potencial MTE definido sobre las variables en el camino que va desde la raíz hasta esa hoja.*

Los árboles mixtos son apropiados para representar potenciales MTE definidos por partes. Cada rama determina una sub-región del espacio donde el potencial está definido, y la función almacenada en la hoja de tal rama es la definición del potencial correspondiente a dicha sub-región.

Ejemplo 9. Consideremos el siguiente potencial MTE, definido para una variable discreta (Y_1) y dos variables continuas (Z_1 y Z_2).

$$\phi(y_1, z_1, z_2) = \begin{cases} 2 + e^{3z_1+z_2} & \text{si } y_1 = 0, 0 < z_1 \leq 1, 0 < z_2 < 2 \\ 1 + e^{z_1+z_2} & \text{si } y_1 = 0, 0 < z_1 \leq 1, 2 \leq z_2 < 3 \\ \frac{1}{4} + e^{2z_1+z_2} & \text{si } y_1 = 0, 1 < z_1 < 2, 0 < z_2 < 2 \\ \frac{1}{2} + 5e^{z_1+2z_2} & \text{si } y_1 = 0, 1 < z_1 < 2, 2 \leq z_2 < 3 \\ 1 + 2e^{2z_1+z_2} & \text{si } y_1 = 1, 0 < z_1 \leq 1, 0 < z_2 < 2 \\ 1 + 2e^{z_1+z_2} & \text{si } y_1 = 1, 0 < z_1 \leq 1, 2 \leq z_2 < 3 \\ \frac{1}{3} + e^{z_1+z_2} & \text{si } y_1 = 1, 1 < z_1 < 2, 0 < z_2 < 2 \\ \frac{1}{2} + e^{z_1-z_2} & \text{si } y_1 = 1, 1 < z_1 < 2, 2 \leq z_2 < 3 \end{cases}$$

Una posible representación de este potencial mediante un árbol mixto podemos verla en la figura 5.2.

5.3. El modelo de regresión Bayes ingenuo basado en MTEs

La propuesta que presentamos en este capítulo consiste, por tanto, en resolver el problema de la regresión en el cual algunas de las variables independientes son discretas o cualitativas usando un modelo tipo Bayes ingenuo en el que la correspondientes distribuciones condicionadas son de tipo MTE. Más concretamente, usaremos un tipo de

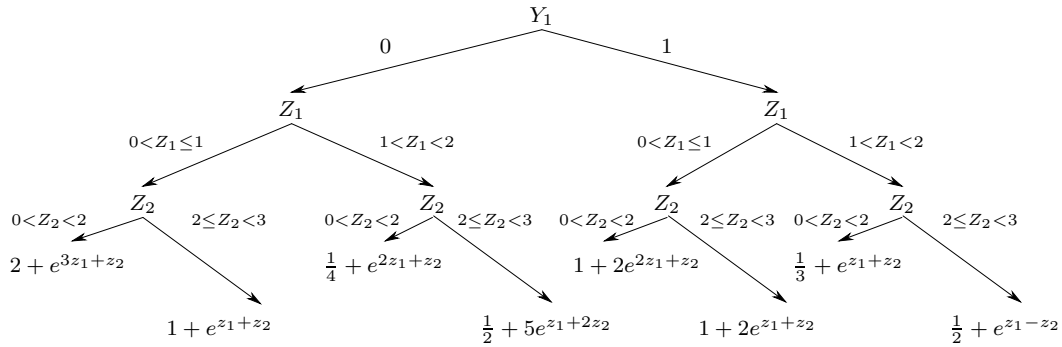


Figura 5.2: Un árbol mixto representando el potencial ϕ del ejemplo 9.

densidad concreto que ofrece un alto poder de ajuste [3], llamada MTE 5-paramétrica, de forma que en cada región en la que se divide el soporte de la variable, ajustaremos una de estas densidades, lo que significa que en cada región (α, β) habría que estimar 5 parámetros a partir de los datos:

$$f(x) = a_0 + a_1 e^{a_2 x} + a_3 e^{a_4 x} \quad \alpha < x < \beta . \quad (5.4)$$

El procedimiento de estimación que hemos seguido es el propuesto en [38], cuyos pasos principales son los siguientes:

1. En primer lugar, se ajusta un kernel gaussiano a partir de los datos.
2. A continuación, el soporte de la variable se divide de acuerdo con los cambios de crecimiento/decrecimiento o concavidad/convexidad del kernel ajustado.
3. Finalmente, en cada división se ajusta un potencial MTE 5-paramétrico al kernel mediante mínimos cuadrados.

Una vez construido el modelo, puede usarse para predecir el valor de la variable

dependiente a partir de las independientes. La predicción se lleva a cabo calculando la distribución de la variable dependiente condicionada a los valores observados de las variables independientes. Esta distribución se calcula mediante *propagación de probabilidad*, para lo que hemos usado la adaptación del algoritmo de Shenoy-Shafer [42] al caso de MTEs propuesta en [41].

Finalmente, la distribución obtenida sirve para proporcionar una predicción numérica de la variable objetivo, ya sea mediante la esperanza, la mediana o la moda. En los casos que hemos analizado, hemos comprobado que la moda siempre proporciona los peores resultados, por lo que a partir de ahora nos centraremos en modelos de predicción usando la mediana y la esperanza.

La esperanza de una variable X cuya densidad está definida como en la ecuación (5.4) se calcula como

$$\begin{aligned}
 E[X] &= \int_{-\infty}^{\infty} x f(x) dx \\
 &= \int_{\alpha}^{\beta} x (a_0 + a_1 e^{a_2 x} + a_3 e^{a_4 x}) dx \\
 &= a_0 \frac{\beta^2 - \alpha^2}{2} + \frac{a_1}{a_2^2} ((a_2 \beta - 1) e^{a_2 \beta} - (a_2 \alpha - 1) e^{a_2 \alpha}) + \\
 &\quad \frac{a_3}{a_4^2} ((a_4 \beta - 1) e^{a_4 \beta} - (a_4 \alpha - 1) e^{a_4 \alpha}) .
 \end{aligned}$$

Si la densidad está definida por partes, la esperanza sería la suma de la expresión anterior en cada una de las partes.

Respecto a la mediana, ésta no puede calcularse al no ser posible invertir la función de distribución correspondiente a la densidad (5.4). Por ello, hemos optado por estimarla mediante el procedimiento de búsqueda descrito en el siguiente algoritmo, que la encuentra con un error inferior a una milésima en términos de probabilidad. El parámetro de entrada es la función de densidad, que puede estar definida en n partes, es decir,

$$f(x) = f_i(x) \quad \alpha_i < x < \beta_i, \quad i = 1, \dots, n,$$

donde cada f_i está definido como en (5.4).

Algoritmo 1. (*Cálculo de la mediana*)

INPUT:

- *La densidad f .*

OUTPUT:

- *Estimación de la mediana.*

1. *found := false*
2. *acum := 0.0*
3. *i := 0*
4. **While** (*(not found) and (i ≤ n)*)
 - a) *m := $\int_{\alpha_i}^{\beta_i} f_i(x)dx$.*
 - b) **If** (*acum + m ≥ 0.5*)
 - *found := true.*
 - c) **Else**
 - *i := i + 1.*
 - *acum := acum + m.*
5. *max := β_i*
6. *min := α_i*
7. *found := false*
8. **While** (*not found*)
 - a) *mid := $\frac{max+min}{2}$*
 - b) *a := $\int_{min}^{mid} f_i(x)dx$*
 - c) *p := acum + a*
 - d) **If** (*$\lfloor 0.5 * 1000 \rfloor == \lfloor p * 1000 \rfloor$*)
 - *found := true.*
 - e) **Else**
 - **If** (*p > 0.5*) *max := mid*
 - **Else** *min := mid.*
9. **Return** *mid.*

5.4. Selección de las variables a incluir en el modelo

Un aspecto importante a tener en cuenta en un modelo de regresión o de clasificación es la selección de las variables que han de incluirse en el modelo. En general, no es cierto que incluir más variables repercuta en una mayor precisión del modelo. Puede ocurrir que algunas variables no sean informativas para la variable dependiente y por lo tanto incluirlas solamente proporcionaría ruido al modelo, además de incrementar la complejidad del mismo en cuanto al número de parámetros a estimar.

Hay tres enfoques al problema de la selección de variables en la situación que nos ocupa:

- El enfoque *filter o filtrado*, que consiste en establecer un ranking de las variables de acuerdo con alguna medida de relevancia respecto a la variable dependiente. Estas medidas de relevancia se llaman normalmente *medidas filter*. A continuación, se fija un umbral de relevancia y todas las variables que no alcancen ese umbral se descartan.
- El enfoque *wrapper o de envoltura*, a diferencia del anterior, consiste en construir distintos modelos con distintos conjuntos de variables independientes. La precisión de cada uno de los modelos se evalúa de alguna forma y el modelo aceptado es el que alcance una mayor precisión.
- El enfoque *filter-wrapper o filtrado-envoltura* es una mezcla de los dos anteriores. En primer lugar, las variables se ordenan de acuerdo con una medida filter y a continuación, se incluyen o excluyen del modelo siguiendo dicho orden, de forma que una variable se incluye si incrementa la precisión del modelo.

Para determinar la precisión de un modelo, hemos seguido un enfoque *train-and-test*, con los siguientes pasos:

1. Disponemos de una base de datos D con información sobre las variables C y X_1, \dots, X_n . Dividimos D en dos partes, D_l y D_t .

2. El modelo se estima usando la base de datos D_l . Normalmente, D_l contiene el 70% de los registros de D , elegidos al azar, mientras que el 30% restante se asigna a D_t . Esta es la proporción que hemos elegido en los estudios de casos que hemos analizado en este capítulo.
3. La precisión del modelo se determina usando la base de datos D_t , midiendo la raíz del error cuadrático medio entre los valores exactos y los proporcionados por el modelo para todos los registros de D_t . Si llamamos c_1, \dots, c_m a los valores de la variable dependiente para los registros de la base de datos D_t y $\hat{c}_1, \dots, \hat{c}_m$ a las correspondientes estimaciones proporcionadas por el modelo, la raíz del error cuadrático medio se calcula como

$$rmse = \sqrt{\frac{1}{m} \sum_{i=1}^m (c_i - \hat{c}_i)^2} . \quad (5.5)$$

El enfoque para la selección de variables que hemos adoptado en este trabajo es el filter-wrapper, usando como medida filter la *información mutua* entre cada variable X_i y C . La información mutua ha sido empleada con éxito en contextos de clasificación [36] y también de regresión [20] bajo la suposición de normalidad. La información mutua entre dos variables X e Y se define como

$$I(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) \log_2 \frac{f_{XY}(x, y)}{f_X(x)f_Y(y)} dydx , \quad (5.6)$$

donde f_{XY} es la densidad conjunta de X e Y , f_X es la densidad marginal de X y f_Y la de Y .

En el caso de densidades MTE, cada densidad viene expresada como en la ecuación (5.4), lo que hace que no podamos calcular el valor de la integral (5.6). Por ello, hemos optado por estimar el valor de la información mutua. El procedimiento de estimación que proponemos está basado en el siguiente teorema.

Teorema 2. Sean X e Y dos variables aleatorias continuas con densidades f_X y f_Y respectivamente, y densidad conjunta f_{XY} . Sea $f_{X|Y}$ la densidad condicionada de X

dada Y . Sea Y_1, \dots, Y_n una muestra extraída de forma independiente a partir de la distribución $f_Y(y)$. Sea X_1, \dots, X_n una muestra tal que cada X_i , $i = 1, \dots, n$ ha sido muestreado a partir de $f_{X|Y}(x|Y_i)$. Entonces,

$$\hat{I}(X, Y) = \frac{1}{n} \sum_{i=1}^n (\log_2 f_{X|Y}(X_i|Y_i) - \log_2 f_X(X_i)) \quad (5.7)$$

es un estimador insesgado de $I(X, Y)$.

Demostración. De acuerdo con la forma en que se obtienen las muestras X_1, \dots, X_n y Y_1, \dots, Y_n , se sigue que la muestra conjunta de puntos $(X_1, Y_1), \dots, (X_n, Y_n)$ tiene distribución de muestreo $f_{XY}(x, y)$. Por lo tanto, si denotamos por $E_{f_{XY}}$ a la esperanza con respecto a la densidad f_{XY} , tenemos que

$$\begin{aligned} E[\hat{I}(X, Y)] &= E_{f_{XY}} \left[\frac{1}{n} \sum_{i=1}^n (\log_2 f_{X|Y}(X_i|Y_i) - \log_2 f_X(X_i)) \right] \\ &= \frac{1}{n} \sum_{i=1}^n E_{f_{XY}} [\log_2 f_{X|Y}(X_i|Y_i) - \log_2 f_X(X_i)] \\ &= \frac{1}{n} \sum_{i=1}^n E_{f_{XY}} \left[\log_2 \frac{f_{X|Y}(X_i|Y_i)}{f_X(X_i)} \right] \\ &= E_{f_{XY}} \left[\log_2 \frac{f_{X|Y}(X|Y)}{f_X(X)} \right] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) \log_2 \frac{f_{X|Y}(x|y)}{f_X(x)} dy dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) \log_2 \frac{f_{XY}(x, y)}{f_X(x) f_Y(y)} dy dx \\ &= I(X, Y) . \end{aligned}$$

□

La consistencia del estimador $\hat{I}(X, Y)$ viene garantizada por la siguiente proposición, siempre que las densidades involucradas sean positivas.

Proposición 3. Sea $\hat{I}_n(X, Y)$ el estimador $\hat{I}(X, Y)$ evaluado para muestras de tamaño n . La sucesión de estimadores $\{\hat{I}_n(X, Y)\}_{n=1}^{\infty}$ es consistente.

Demostración. Basta con demostrar que

$$(i) \lim_{n \rightarrow \infty} E[\hat{I}_n(X, Y)] = I(X, Y) \text{ y}$$

$$(ii) \lim_{n \rightarrow \infty} \text{Var}(\hat{I}_n(X, Y)) = 0.$$

La demostración de (i) es trivial, dado que según el teorema 2, $E[\hat{I}_n(X, Y)] = I(X, Y)$ para todo $n > 0$ y por lo tanto el límite también es igual a $I(X, Y)$.

De cara a demostrar (ii), necesitamos la expresión de la varianza del estimador.

$$\begin{aligned} \text{Var}(\hat{I}_n(X, Y)) &= \text{Var} \left(\frac{1}{n} \sum_{i=1}^n (\log_2 f_{X|Y}(X_i|Y_i) - \log_2 f_X(X_i)) \right) \\ &= \frac{1}{n} \text{Var} (\log_2 f_{X|Y}(X|Y) - \log_2 f_X(X)) \end{aligned}$$

donde $\text{Var} (\log_2 f_{X|Y}(X|Y) - \log_2 f_X(X))$ no depende de n y es finita siempre que las densidades $f_{X|Y}$ y f_X sean positivas. Por tanto, podemos concluir que

$$\lim_{n \rightarrow \infty} \text{Var}(\hat{I}_n(X, Y)) = 0 .$$

□

5.4.1. Algoritmo para la construcción del modelo con selección de variables.

En este punto, estamos en condiciones de presentar un algoritmo detallado para la construcción del modelo de regresión tipo Bayes ingenuo basado en MTEs y con selección de variables. Llamaremos a este modelo *predictor Bayes ingenuo selectivo*.

Algoritmo 2 (Predictor Bayes ingenuo selectivo).

INPUT:

- *La variable dependiente C .*
- *Las variables independientes X_1, \dots, X_n .*
- *Una base de datos D con variables X_1, \dots, X_n, C .*

OUTPUT:

- *Un predictor Bayes ingenuo selectivo para la variable C .*

1. For $i := 1$ to n

- *Calcula $\hat{I}(X_i, C)$.*

2. Sea $X_{(1)}, \dots, X_{(n)}$ un orden decreciente de las variables independientes de acuerdo con $\hat{I}(X_{(i)}, C)$.

3. Divide la base de datos D en dos partes, una para estimar el modelo (D_l) y otra para determinar la precisión del mismo (D_t).

4. Construye un predictor Bayes ingenuo, M , para las variables C y $X_{(1)}$ como sigue:

- a) *Estima una densidad marginal MTE para C , f_C , a partir de la base de datos D_l usando el método propuesto en [38].*
- b) *Estima una densidad MTE condicionada para $X_{(1)}$ dada C , $f_{X_{(1)}|C}$, a partir de la base de datos D_l .*
- c) *Sea $rmse(M)$ la precisión estimada del modelo M usando la base de datos D_t , de acuerdo con la fórmula (5.5).*

5. For $i := 2$ to n

- a) *Sea M_1 el predictor Bayes ingenuo obtenido a partir de M añadiendo la variable $X_{(i)}$, es decir, estimando una densidad condicionada para $X_{(i)}$ dada C , $f_{X_{(i)}|C}$, a partir de la base de datos D_l .*
- b) *Sea $rmse(M_1)$ la precisión estimada del modelo M_1 usando la base de datos D_t , de acuerdo con la fórmula (5.5).*
- c) **If** ($rmse(M_1) \leq rmse(M)$)
 - $M := M_1$.

6. **Return** (M)

5.5. Estudio de casos

Hemos empleado el predictor Bayes ingenuo selectivo en distintos problemas prácticos relacionados con la estadística universitaria. Concretamente, tratamos de predecir el valor de distintos indicadores, referidos a titulaciones, de los introducidos en la sección 1.4 del capítulo 1 así como del número de alumnos en una asignatura dada.

Hemos considerado 6 problemas:

- **Problema 1:** Predicción del valor de la *tasa de rendimiento*.
- **Problema 2:** Predicción del valor de la *tasa de éxito*.
- **Problema 3:** Predicción del valor de la *tasa de graduación progresiva*.
- **Problema 4:** Predicción del valor de la *tasa de graduación*.
- **Problema 5:** Predicción del valor del *coeficiente de duración media*.
- **Problema 6:** Predicción del número de alumnos en una asignatura concreta.

En los problemas 1 al 5 hemos considerado una base de datos con los valores de las siguientes variables referentes a titulaciones en la Universidad de Almería entre los cursos 2000-01 y 2003-04.

Aunque los indicadores que se relacionan a continuación han sido definidos en la sección 2.3, recordamos su significado para simplificar la lectura.

- **Titulación:** código de la titulación.
- **TasaOpt:** tasa de optatividad, definida como ratio entre el número total de créditos optativos y de libre configuración que ha de cursar el alumnado a lo largo de la carrera y el número total de créditos requeridos.
- **OfertaOpt:** número de créditos ofertados en optatividad dividido por el número de créditos optativos a cursar por el alumno.

Variable	Información Mutua
OfertaOpt	0.1998
Prt	0.1791
GruposP	0.1361
TasaOpt	0.1357
Dedicación	0.1119
TasaAbandono	0.1033
Ratio_A-P	0.0692
Titulación	0.0428
GruposG	0.0235
Doct	0.0169
NotaAcceso	0

Tabla 5.1: Información mutua estimada entre cada variable predictora y la variable dependiente para el problema 1 (predicción de la tasa de rendimiento)

- **Prt**: ratio entre el número total de créditos prácticos requeridos y el número total de créditos incluidos en el plan de estudios.
- **GruposP**: ratio entre el número de grupos de teoría con no más de 20 alumnos entre el número total de grupos teóricos.
- **GruposG**: ratio entre el número de grupos de teoría con no menos de 80 alumnos y el número total de grupos teóricos.
- **Dedicación**: número de créditos matriculados por los estudiantes dividido por el número de estudiantes.
- **TasaAbandono**: tasa de estudiantes que abandonan la titulación.
- **Ratio_A-P**: número de estudiantes por profesor.
- **Doct**: fracción de créditos impartidos por doctores.
- **NotaAcceso**: media de las notas de acceso de los estudiantes en una titulación.

Variable	Información Mutua
Prt	0.1400
GruposP	0.1026
Dedicación	0.0939
TasaOpt	0.0583
OfertaOpt	0.0544
Ratio_A-P	0.0497
Titulación	0.0262
TasaAbandono	0.0253
GruposG	0.0126
Doct	0.0098
NotaAcceso	0

Tabla 5.2: Información mutua estimada entre cada variable predictora y la variable dependiente para el problema 2 (predicción de la tasa de éxito)

El problema 6 consiste en estimar el número de estudiantes en una asignatura. En este caso, hemos usado una base de datos con 354 registros referentes a las variables que a continuación se detallan, acerca de las asignaturas troncales y obligatorias ofertadas en la Universidad de Almería en los cursos desde 2001-02 hasta 2005-06. Por tanto, la predicción del número de alumnos se referirá al curso 2005/2006. Las variables empleadas son:

- **Titulación:** código de la titulación.
- **Ciclo:** ciclo en el que se oferta la asignatura.
- **CódigoAsignatura.**
- **Curso:** curso en el que se imparte la asignatura.
- **EXX:** número de estudiantes en la asignatura en cuestión en el año XX, desde 01 hasta 04.
- **RXX:** tasa de rendimiento para la asignatura en cuestión en el año XX, desde 01 hasta 04.

Variable	Información Mutua
OfertaOpt	0.1317
Prt	0.0920
GruposP	0.0699
Dedicación	0.0676
Titulación	0.0643
TasaOpt	0.0534
Ratio_A-P	0.0410
Doct	0.0340
TasaAbandono	0.0281
GruposG	0.0197
NotaAcceso	0

Tabla 5.3: Información mutua estimada entre cada variable predictora y la variable dependiente para el problema 3 (predicción de la tasa de graduación progresiva)

En todos los casos, hemos usado los siguientes modelos:

- **NB(media)**: Predictor Bayes ingenuo (naive Bayes) incluyendo todas las variables y realizando la predicción con la media de la distribución a posteriori.
- **NB(media)**: Predictor Bayes ingenuo (naive Bayes) incluyendo todas las variables y realizando la predicción con la mediana de la distribución a posteriori.
- **SNB(media)**: Predictor Bayes ingenuo selectivo obtenido mediante el algoritmo 2 y realizando la predicción con la media de la distribución a posteriori.
- **SNB(media)**: Predictor Bayes ingenuo selectivo obtenido mediante el algoritmo 2 y realizando la predicción con la mediana de la distribución a posteriori.
- **LM**: Modelo de regresión lineal incluyendo todas las variables, y tomando las discretas como si fueran continuas.
- **SLM**: Modelo lineal incluyendo sólo las variables empleadas por el mejor de el modelo Bayes ingenuo selectivo (es decir, SNB(media) o SNB(media)).

Variable	Información Mutua
Doct	0.2088
GruposP	0.1277
TasaAbandono	0.1114
Prt	0.0951
Dedicación	0.0942
GruposG	0.0665
TasaOpt	0.0546
Titulación	0.0261
Ratio_A-P	0.0126
OfertaOpt	0.0091
NotaAcceso	0

Tabla 5.4: Información mutua estimada entre cada variable predictora y la variable dependiente para el problema 4 (predicción de la tasa de graduación)

Las informaciones mutuas estimadas para las variables de los distintos problemas estudiados pueden verse en las tablas 5.1 hasta 5.6. Los modelos construidos se muestran en las figuras 5.3 hasta 5.11, donde cada nodo de la red representada contiene la distribución marginal a priori de la correspondiente variable. La forma de las densidades obtenidas muestra la capacidad de ajuste de las MTEs en distintas situaciones, por la variedad de formas que pueden adoptar. En el caso en que el predictor selectivo contenga un conjunto de variables distinto según se use la media o la mediana para predecir, se muestran ambos casos.

5.5.1. Discusión de los resultados

En cinco de los seis problemas analizados, el predictor Bayes ingenuo selectivo proporciona mejores resultados que el modelo lineal. Sólo en el caso 6, de predicción del número de estudiantes, el modelo lineal es claramente mejor. Esto indica que la situación representada por el problema 6 es la que más se acerca a la linealidad. Las redes bayesianas han mostrado su mayor adecuación a situaciones de falta de linealidad.

Variable	Información Mutua
OfertaOpt	0.1743
TasaAbandono	0.1647
Prt	0.1481
NotaAcceso	0.1346
Ratio_A-P	0.1067
GruposP	0.1045
TasaOpt	0.1040
GruposG	0.0681
Dedicación	0.0593
Titulación	0.0557
Doct	0.0141

Tabla 5.5: Información mutua estimada entre cada variable predictora y la variable dependiente para el problema 5 (predicción del coef. de duración media)

En cuanto a la selección de variables, los resultados indican que siempre proporciona importantes mejoras, lo que apoya la corrección del procedimiento de selección diseñado.

Respecto a predecir con la media o con la mediana, en cuatro de los seis problemas la mediana ha resultado mejor cuando se habla del predictor selectivo, pero con pequeñas diferencias en general. No hemos proporcionado resultados para la moda pues en todos los experimentos piloto que realizamos, éstos estaban muy alejados de los que ofrecían tanto la media como la mediana.

Finalmente cabe indicar que un valor añadido tanto del predictor NB como del SNB respecto al modelo lineal, es que los dos primeros no sólo proporcionan resultados numéricos para la predicción, sino que también dan como resultado la distribución a posteriori de la variable dependiente. Esto es importante porque permite realizar otro tipo de inferencias como por ejemplo responder a preguntas del tipo *¿cuál es la probabilidad de que el número de estudiantes esté entre 100 y 150?*

Variable	Información Mutua
E04	0.8338
E03	0.6853
E02	0.5795
E01	0.4953
Titulación	0.0865
Ciclo	0.0630
R01	0.0526
R04	0.0520
CódigoAsig	0.0516
R03	0.0470
R02	0.0450
Curso	0.0257

Tabla 5.6: Información mutua estimada entre cada variable predictora y la variable dependiente para el problema 6 (predicción del número de estudiantes)

5.6. Conclusiones

En este capítulo hemos presentado un marco para afrontar problemas de regresión donde algunas de las variables independientes son discretas y también cuando todas son continuas pero la distribución conjunta no es normal multivariante. El enfoque propuesto está basado en el empleo de redes bayesianas y distribuciones de tipo MTE como modelo probabilístico subyacente. También hemos propuesto un esquema de selección de variables basado en la información mutua, con una estrategia filter-wrapper.

El modelo diseñado ha sido aplicado a seis problemas reales de predicción en el marco de la estadística universitaria, mostrando un comportamiento razonablemente bueno salvo en uno de los casos, en el que el modelo lineal se comporta mejor.

Método	rmse	Núm. variables
NB(media)	0.0884	11
NB(media)	0.0921	11
SNB(media)	0.0818	9
SNB(media)	0.0814	6
LM	0.1154	11
SLM	0.1036	6

Tabla 5.7: Resultados para el problema 1 (tasa de rendimiento)

Método	rmse	Núm. variables
NB(media)	0.0462	11
NB(media)	0.0471	11
SNB(media)	0.0383	8
SNB(media)	0.0381	8
LM	0.0476	11
SLM	0.0476	8

Tabla 5.8: Resultados para el problema 2 (tasa de éxito)

Método	rmse	Núm. variables
NB(media)	0.0815	11
NB(media)	0.0908	11
SNB(media)	0.0657	5
SNB(media)	0.0671	7
LM	0.0668	11
SLM	0.0666	5

Tabla 5.9: Resultados para el problema 3 (tasa de graduación progresiva)

Método	rmse	Núm. variables
NB(media)	10.2648	11
NB(media)	10.6803	11
SNB(media)	7.6028	4
SNB(media)	7.8870	5
LM	10.2380	11
SLM	9.8816	4

Tabla 5.10: Resultados para el problema 4 (tasa de graduación)

Método	rmse	Núm. variables
NB(media)	0.4328	11
NB(media)	0.4265	11
SNB(media)	0.2722	7
SNB(media)	0.2568	7
LM	0.2972	11
SLM	0.2831	7

Tabla 5.11: Resultados para el problema 5 (coef. duración media)

Método	rmse	Núm. variables
NB(mean)	31.1731	12
NB(median)	31.7893	12
SNB(mean)	23.6530	7
SNB(median)	23.6138	7
LM	16.4054	12
SLM	18.0076	7

Tabla 5.12: Resultados para el problema 6 (número de alumnos)

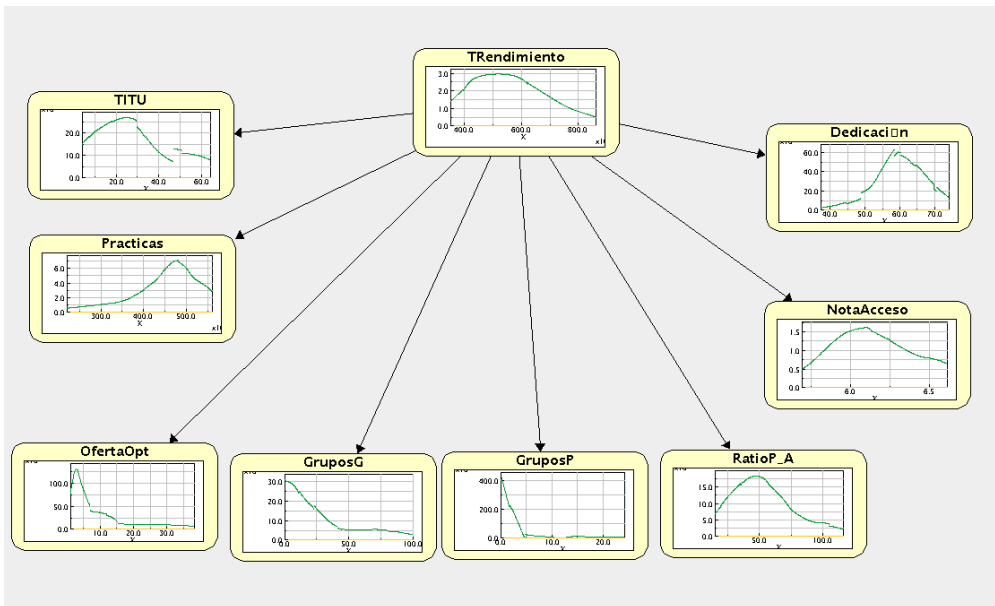


Figura 5.3: Predictor SNB(media) para el problema 1.

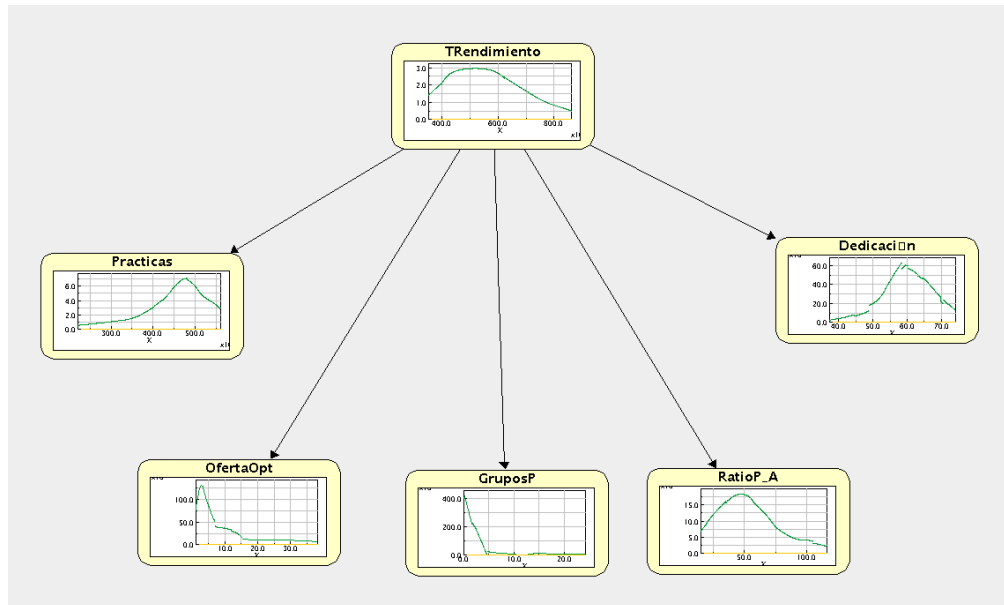


Figura 5.4: Predictor SNB(media) para el problema 1.

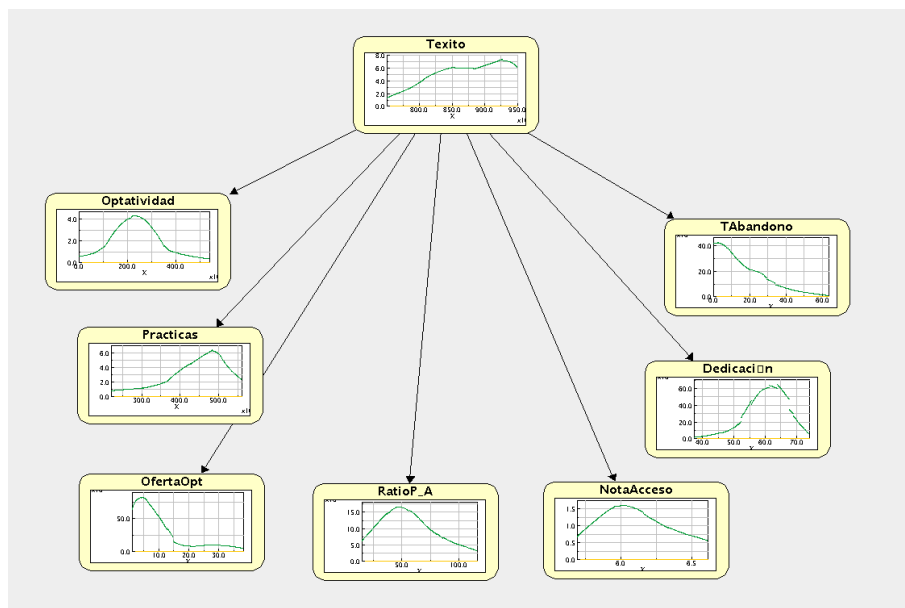


Figura 5.5: Predictor SNB(media y mediana) para el problema 2.

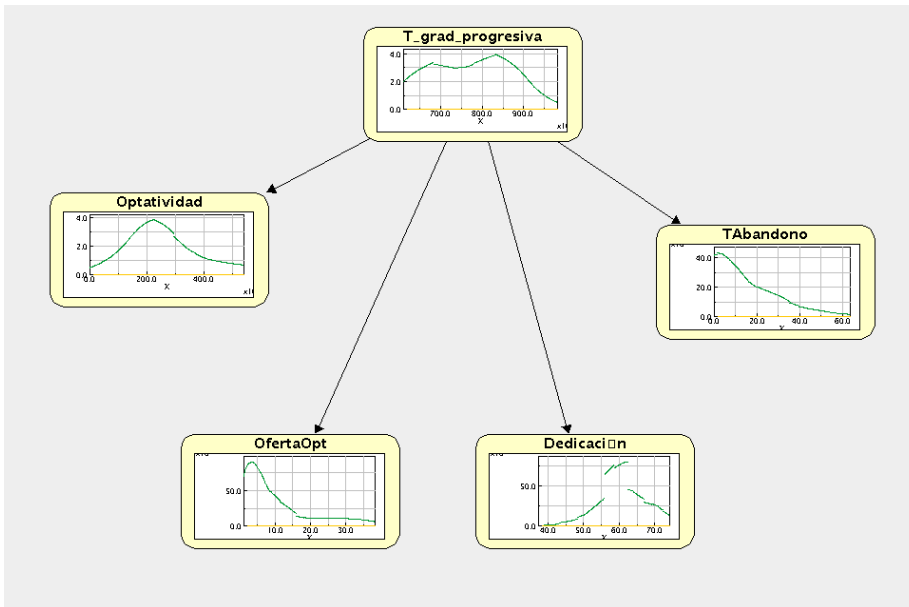


Figura 5.6: Predictor SNB(media) para el problema 3.

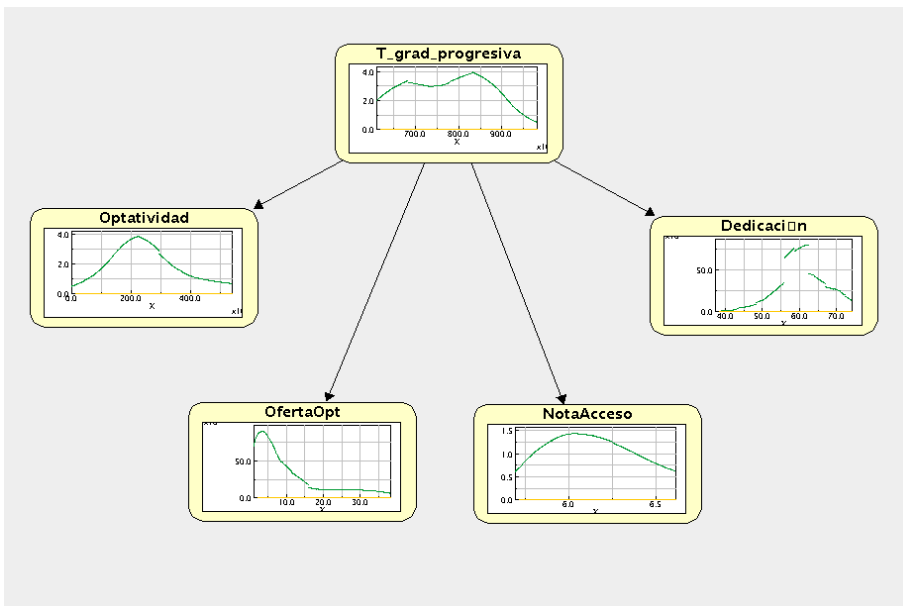


Figura 5.7: Predictor SNB(mediana) para el problema 3.

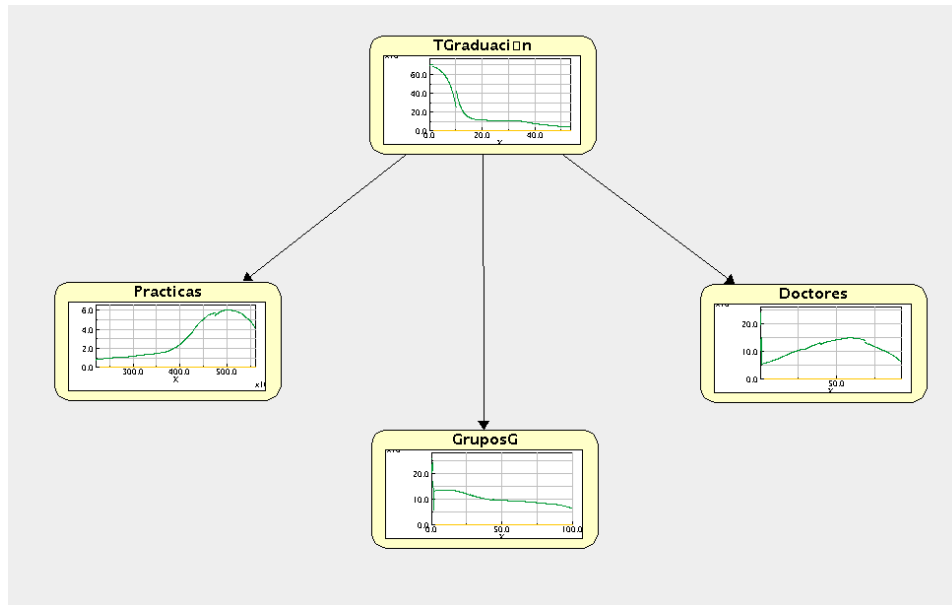


Figura 5.8: Predictor SNB(media) para el problema 4.

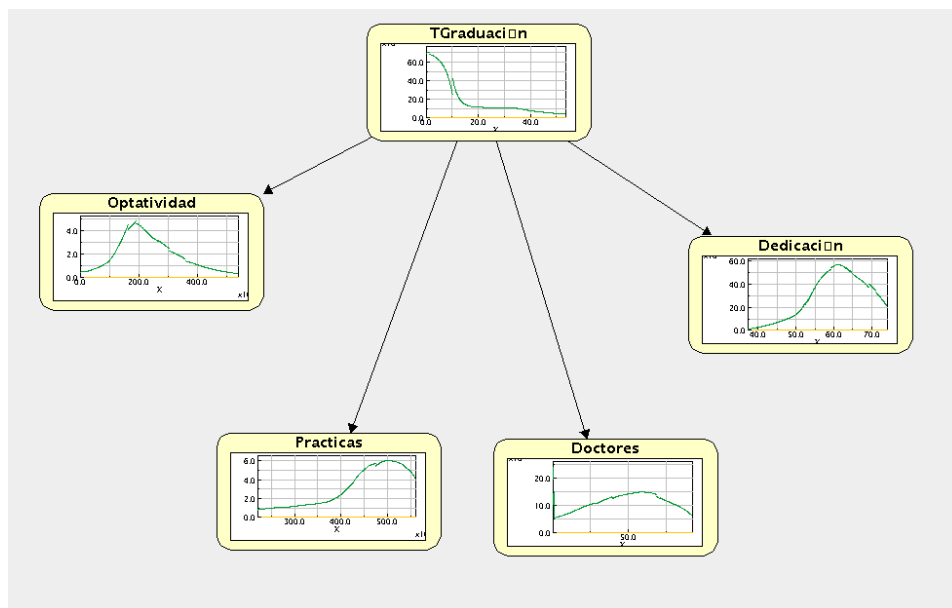


Figura 5.9: Predictor SNB(media) para el problema 4.

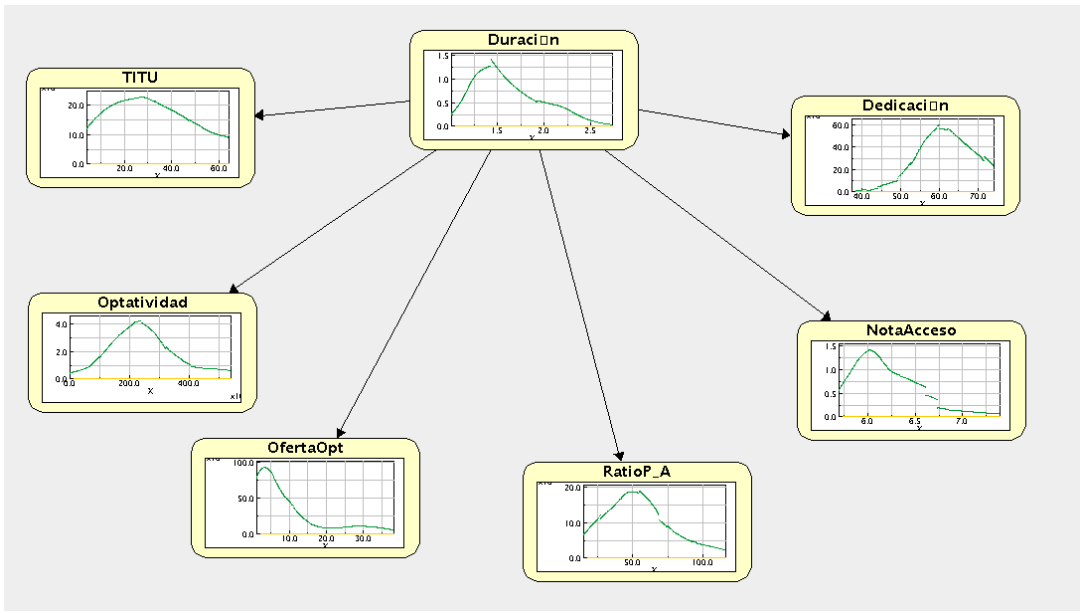


Figura 5.10: Predictor SNB(media y mediana) para el problema 5.

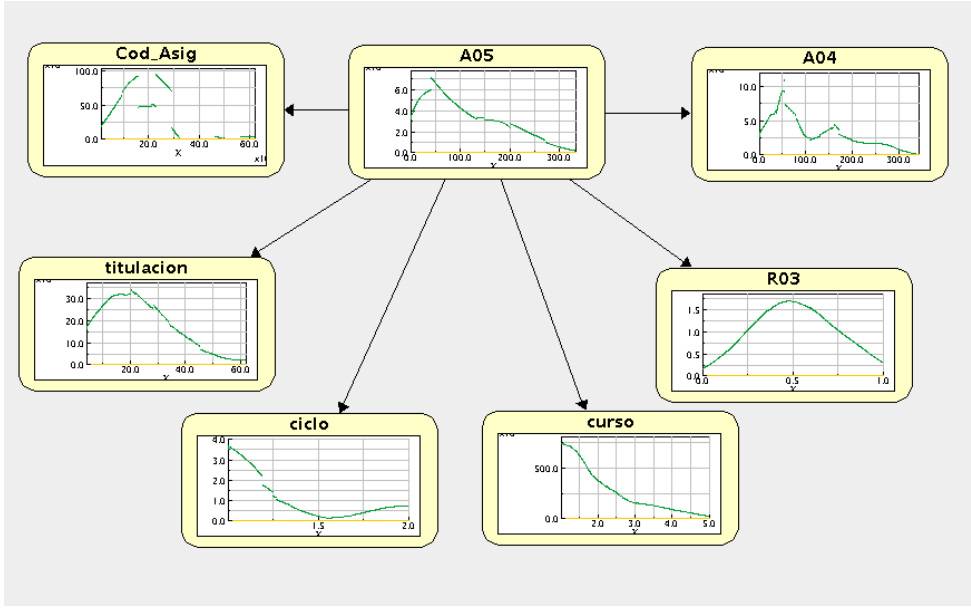


Figura 5.11: Predictor SNB(media y mediana) para el problema 6.

Capítulo 6

CONCLUSIONES Y LÍNEAS ABIERTAS

En esta memoria hemos abordado el estudio de los indicadores universitarios, que en los últimos años se ha convertido en una herramienta fundamental en la toma de decisiones en la administración de las instituciones universitarias.

La reducción de los recursos públicos para la financiación de las universidades motivan la introducción de sistemas de financiación vinculada a la consecución de resultados. En este contexto, los indicadores de rendimiento se revelan como el soporte más adecuado para estos nuevos mecanismos de asignación de recursos. Sin embargo, la construcción de un sistema de indicadores adecuado para medir el rendimiento de una institución como la universitaria, plantea una problemática que es tratada detalladamente a lo largo de los primeros capítulos.

Una vez establecido el marco para el uso de los indicadores de rendimiento, se impone el problema de la elección de un mecanismo que ayude al análisis de los datos generados por la universidad y apoye la toma de decisiones por parte de los directivos de ésta. Proponemos las redes bayesianas porque ofrecen una herramienta clara y de fácil manejo para usuarios no versados en técnicas estadísticas: como modelo gráfico, facilita la comprensión del problema, presentando, además, la estructura de dependen-

cias entre las variables. Por otro lado, las técnicas de propagación ofrecen la posibilidad de estudiar los efectos que tendrían las posibles líneas de actuación sobre las variables de interés.

Nos hemos centrado en el problema práctico del estudio del rendimiento en las asignaturas y titulaciones, ya que es uno de los objetivos a cumplir en el contrato-programa suscrito por la Universidad de Almería, y del cual depende el 10% de su financiación. En el caso del estudio por asignatura, hemos considerado las variables que influyen en la tasa de rendimiento, analizando su efecto y estableciendo que la variable con mayor ascendente sobre dicha tasa es el tamaño de los grupos de teoría.

Realizando un estudio análogo para el rendimiento académico en las titulaciones, se deriva una serie de medidas que la dirección de la Universidad podría llevar a cabo para la mejora de las tasas de rendimiento, de graduación, de duración media o de abandono. Estas medidas afectan tanto al plan de estudios (aumento del peso práctico o de carga optativa), como a la ordenación docente, disminuyendo el tamaño de los grupos teóricos.

Por último, abordamos la predicción de magnitudes en estadística universitaria como herramienta para la toma de decisiones. La casuística de los datos universitarios (heterogeneidad, imposibilidad de realizar un estudio de la evoluciones en el tiempo debido a la escasez de datos) motiva la propuesta de un modelo de regresión basado en redes bayesianas.

En este marco, hemos desarrollado un modelo que se fundamenta en un tipo de red bayesiana conocida como *Bayes ingenuo*. El problema del tratamiento conjunto de variables discretas y continuas lo hemos resuelto utilizando MTEs, realizando una selección de variables basado en la información mutua con una estrategia filter-wrapper.

Aplicando el *predictor Bayes ingenuo selectivo* a seis problemas reales de predicción en estadística universitaria, se han obtenido mejores resultados que el modelo de regresión lineal en 5 de los seis problemas.

6.1. Líneas de trabajo futuro

En el apartado de indicadores, nos planteamos como objetivo inmediato la definición de indicadores globales que de alguna forma resuman el comportamiento de otro conjunto de indicadores. De esta forma, sería posible monitorizar el grado de cumplimiento de unos objetivos que se evalúa mediante una serie de indicadores. Pensamos en situaciones en las que haya que determinar si la evolución de un cuadro de mando institucional ha sido satisfactoria, cuando posiblemente algunos de los indicadores que lo integran no hayan evolucionado de forma positiva.

En el campo de la predicción, creemos que el modelo de regresión Bayes ingenuo selectivo ofrece amplias posibilidades de aplicación. Por ello, pretendemos abordar el estudio de problemas prácticos donde tradicionalmente se ha empleado el modelo lineal.

Nos proponemos profundizar en la mejora del modelo de regresión Bayes ingenuo selectivo. Por un lado, perfeccionando el procedimiento de selección de variables, empleando otras medidas distintas de la información mutua y probando procedimientos filter-wrapper más elaborados que el empleado en esta memoria. Por otro lado, la suposición de independencia condicional de las variables independientes dada la variable dependiente puede relajarse mediante la consideración de estructuras de red más complejas.

Bibliografía

- [1] En T.Wyatt, editor, *Education indicators: An overview*. OCDE/INES, 1991.
- [2] R. Alemany. Indicadores de calidad docente: ¿café para todos?. En J. Vidal García, editor, *Indicadores en la Universidad: información y decisiones*, páginas 301–310. Consejo de Universidades. Secretaría General, 1999.
- [3] B. Cobb, R. Rumí y A. Salmerón. Modeling conditional distributions of continuous variables in Bayesian networks. *Lecture Notes in Computer Science*, 3646:36–45, 2005.
- [4] B. Cobb, P.P. Shenoy y R. Rumí. Approximating probability density functions with mixtures of truncated exponentials. *Statistics and Computing*, In press, 2006.
- [5] S. Cuenin. The use of performance indicators in universities: an international survey. *International journal of management in Higher Education*, 2:117–139, 1987.
- [6] S. Cuenin. Un examen de las políticas europeas en materia de evaluación de la enseñanza superior. En M. De Miguel, J.G. Mora y S. Rodríguez, editores, *La evaluación de las instituciones universitarias*, páginas 371–397. Consejo de Universidades. Secretaría General, 1991.
- [7] L.M. de Campos. Aprendizaje automático de modelos gráficos. En J.A. Gámez y J.M. Puerta, editores, *Sistemas expertos probabilísticos*, páginas 113–140. Ediciones de la Universidad de Castilla-La Mancha, 1998.

- [8] Consejo de Coordinación Universitaria. Borrador del sistema de indicadores del sistema universitario público español. <http://www.mec.es/educa/ccuniv/html/indicadores/indicado.pdf>, 2000. Consejo de Coordinación Universitaria, Ministerio de Educación y Ciencia.
- [9] J. de Miguel, J. Caïs y E. Vaquera. Modelo de indicadores de calidad de las universidades. *Gestión y análisis de políticas públicas*, 16:29–62, 1999.
- [10] M. De Miguel Díaz. Utilización de indicadores en la evaluación de la docencia universitaria. En M. de Miguel, J. Mora y S. Rodríguez, editores, *La evaluación de las instituciones universitarias*, páginas 341–370. Consejo de Universidades. Secretaría General, 1991.
- [11] M. De Miguel Díaz. Indicadores de rendimiento y evaluación de programas. En R. Pérez Juste y otros, editores, *Evaluación de programas y centros educativos*, páginas 169–192. Cincel, 1995.
- [12] F. Dochy, M. Segers y W. Wijnen. Selecting performance indicators. a proposal as a result of research. En F. Goedegebuure, F. Maasen y D. Westerheijden, editores, *Peer review and performance indicators*, páginas 135–153. Lemma B.V., 1990.
- [13] F. Dochy, M. Segers y W. Wijnen. Selección de indicadores de rendimiento. una propuesta como resultado de una investigación. En M. de Miguel, J.G. Mora y S. Rodríguez, editores, *La evaluación de las instituciones universitarias*, páginas 317–339. Consejo de Universidades. Secretaría General, 1991.
- [14] R.O. Duda, P.E. Hart y D.G. Stork. *Pattern classification*. Wiley Interscience, 2001.
- [15] Elvira Consortium. Elvira: An environment for creating and using probabilistic graphical models. En J.A. Gámez y A. Salmerón, editores, *Proceedings of the First European Workshop on Probabilistic Graphical Models*, páginas 222–230, 2002.
- [16] E. Frackmann. Lecciones que deben aprenderse de una década de discusiones sobre indicadores de rendimiento. En M. de Miguel, J.G. Mora y S. Rodríguez, editores,

- La evaluación de las instituciones universitarias*, páginas 399–421. Consejo de Universidades. Secretaría General, 1991.
- [17] E. Frank, L. Trigg, G. Holmes y I.H. Witten. Technical note: Naive Bayes for regression. *Machine Learning*, 41:5–25, 2000.
- [18] N. Friedman, D. Geiger y M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.
- [19] J.A. Gámez. Abducción en modelos gráficos. En J.A. Gámez y J.M. Puerta, editores, *Sistemas expertos probabilísticos*, páginas 89–111. Ediciones de la Universidad de Castilla-La Mancha, 1998.
- [20] J.A. Gámez y A. Salmerón. Predicción del valor genético en ovejas de raza manchega usando técnicas de aprendizaje automático. En *Actas de las VI Jornadas de Transferencia de Tecnología en Inteligencia Artificial*, páginas 71–80. Paraninfo, 2005.
- [21] J. Hernández. *La Universidad española en cifras*. CRUE, 2004.
- [22] J. Hernández, M.J. Ramírez y C. Ferri. *Introducción a la minería de datos*. Pearson, 2004.
- [23] L.D. Hernández. Algoritmos de propagación i: métodos exactos. En J.A. Gámez y J.M. Puerta, editores, *Sistemas expertos probabilísticos*, páginas 41–64. Ediciones de la Universidad de Castilla-La Mancha, 1998.
- [24] H.Jacobson. Framework for evaluation: Indicators of effort, performance, effects. En H.Jacobson, editor, *Evaluating advancement programs, new directions for institutional advancement*. Jossey-Bass, 1978.
- [25] A. Jarrat. Efficiency studies in universities, 1985. Committee of Vice Chancellors and Principals. London.
- [26] F.V. Jensen. *An introduction to Bayesian networks*. UCL Press, 1996.

- [27] J.F.Huete. Sistemas expertos probabilísticos: modelos gráficos. En J.A. Gámez y J.M. Puerta, editores, *Sistemas expertos probabilísticos*, páginas 1–40. Ediciones de la Universidad de Castilla-La Mancha, 1998.
- [28] S. Kagan y M.Smith. Indicators of educational quality. *Educational Leadership*, pages 16–32, 1985.
- [29] S. Moral, R. Rumí y A. Salmerón. Mixtures of truncated exponentials in hybrid Bayesian networks. En *Lecture Notes in Artificial Intelligence*, volume 2143, páginas 135–143, 2001.
- [30] M. Morales, C. Rodríguez y A. Salmerón. Estudio de dependencias de indicadores de rendimiento del alumnado universitario mediante redes bayesianas. En *Actas de las VI Jornadas de Transferencia de Tecnología en Inteligencia Artificial*, páginas 29–36, 2005.
- [31] M. Morales, C. Rodríguez y A. Salmerón. Selective naive Bayes predictor with mixtures of truncated exponentials. En *Proceedings of the International Conference on Mathematical and Statistical Modeling*, 2006.
- [32] M. Morales y A. Salmerón. Análisis del alumnado de la Universidad de Almería mediante redes bayesianas. En *Actas del 27 Congreso Nacional de Estadística e I.O.*, páginas 3413–3436, 2003.
- [33] M.S.Smith. Educational indicators. *Phi Delta Kappan*, 7:487–491, 1988.
- [34] J. Oakes. *Educational indicators: A guide for policymakers*. Rutgers University, 1986.
- [35] J. Pearl. *Probabilistic reasoning in intelligent systems*. Morgan-Kaufmann (San Mateo), 1988.
- [36] A. Pérez, P. Larrañaga y I. Inza. Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes. *International Journal of Approximate Reasoning*, In press, 2006.

- [37] E. Villareal Rodríguez. La utilización de indicadores de rendimiento en la financiación de la educación superior. En J. Vidal García, editor, *Indicadores en la Universidad: información y decisiones*, páginas 65–79. Consejo de Universidades. Secretaría General, 1999.
- [38] V. Romero, R. Rumí y A. Salmerón. Learning hybrid Bayesian networks using mixtures of truncated exponentials. *International Journal of Approximate Reasoning*, 42:54–68, 2006.
- [39] J.G. Mora Ruiz. *Calidad y rendimiento en las instituciones universitarias*. Consejo de Universidades. Secretaría General, Madrid, España, 1991.
- [40] J.G. Mora Ruiz. Indicadores para la información, la gestión y la financiación de las universidades. En UTREF, editor, *Nuevas miradas sobre la Universidad*, páginas 15–34, 2002.
- [41] R. Rumí y A. Salmerón. Penniless propagation with mixtures of truncated exponentials. *Lecture Notes in Computer Science*, 3571:39–50, 2005.
- [42] P.P. Shenoy y G. Shafer. Axioms for probability and belief function propagation. En R.D. Shachter, T.S. Levitt, J.F. Lemmer, and L.N. Kanal, editores, *Uncertainty in Artificial Intelligence 4*, páginas 169–198. North Holland, Amsterdam, 1990.
- [43] J. Sizer. Assessing institutional performance: An overview. *International Journal of Institutional Management in Higher Education*, 1:49–77, 1979.
- [44] J. Vidal. Indicadores de rendimiento para las universidades españolas: necesidad y disponibilidad. En J. Vidal García, editor, *Indicadores en la Universidad: información y decisiones*, páginas 65–79. Consejo de Universidades. Secretaría General, 1999.
- [45] M. Yorke. Siamese twins? performance indicators in the service of accountability and enhancement. *Quality in Higher Education*, 1:13–30, 1995.