Estudio de Dependencias de Indicadores de Rendimiento del Alumnado Universitario mediante Redes Bayesianas*

María Morales Carmelo Rodríguez Antonio Salmerón

Maria.Morales@ual.es crt@ual.es Antonio.Salmeron@ual.es

Dept. de Estadística y Matemática Aplicada

Universidad de Almería

Ctra. Sacramento s/n

04120 Almería

Resumen

En los últimos años ha cobrado gran importancia el uso de indicadores para describir el perfil de las Universidades españolas en términos tanto académicos como investigadores y económicos. Estos indicadores son utilizados para tomar decisiones de gran importancia, llegando a afectar incluso a aspectos de financiación. Sin embargo, el número de indicadores a veces es excesivo, lo que aumenta el riesgo de redundancia y disfuncionalidad. En este trabajo presentamos un estudio mediante redes bayesianas de un grupo de indicadores de rendimiento, particularizado al caso de la Universidad de Almería en el curso 2003/2004.

1. Introducción

Cada vez está más generalizado el uso de indicadores en todas las facetas relacionadas con la gestión universitaria [3]. Estos indicadores son utilizados para tomar decisiones de gran importancia, llegando a afectar incluso a aspectos de financiación. Sin embargo, el número de indicadores a veces es excesivo, lo que aumenta el riesgo de redundancia y disfuncionalidad. Esta situación aconsejaba la realización de un estudio basado en análisis de situaciones reales que de alguna forma arrojara luz sobre el comportamiento de algunos indicadores. El problema que presenta un análisis estadístico tradicional para este tipo de situaciones radica en el alto número de variables y de datos a manejar, así como en la heterogeneidad de las variables. En este sentido, las redes bayesianas [6] 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 incluso llevar a una mejor comprensión del mismo. Un ejemplo es el estudio sobre el alumnado de la Universidad de Almería contenido en [5].

En este trabajo presentamos un estudio mediante redes bayesianas de un grupo de indicadores de rendimiento, particularizado al caso de la Universidad de Almería en el curso 2003/2004. Para ello hemos considerado distintos escenarios desde el punto de vista de las asignaturas.

El resto del trabajo está organizado como sigue. En la sección 2 revisamos los conceptos básicos sobre redes bayesianas y su construcción a partir de datos. La sección 3 describe los conjuntos de datos empleados y los cuatro escenarios de asignaturas considerados en el análisis que se lleva a cabo en la sección 4. El articulo finaliza con las conclusiones en la sección 5.

*Trabajo subvencionado por el Ministerio de Educación y Ciencia y por fondos FEDER, bajo el proyecto TIN2004-06204-C03-01

2. Modelos basados en redes bayesianas

Una *red bayesiana* es un grafo dirigido acíclico donde cada nodo representa una variable aleatoria y tal que asociada a cada nodo hay una distribución de probabilidad condicionada a sus padres en el grafo.

Una red bayesiana representa una distribución de probabilidad multivariante, de manera que las relaciones de independencia entre las variables que la forman quedan identificadas de forma gráfica mediante el concepto de *d*-separación [6].

Definición 1. *Dos variables A y B en una red bayesiana se dice que están* **d-separadas** *si todos los caminos entre A y B son como los que aparecen en la figura 1. Donde en los dos primeros casos la variable C está observada y en el tercer caso ni C ni ninguno de sus descendientes lo está. Se dice además que C d-separa a A y B.*

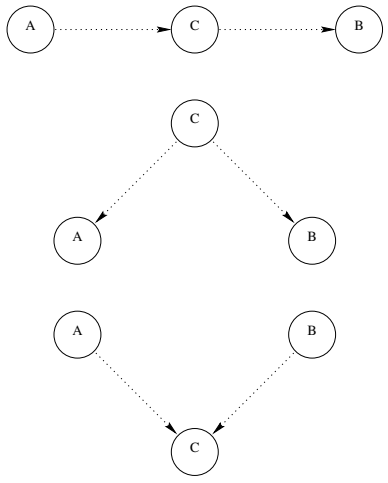

Figura 1: Caracterización gráfica del concepro de d-separación.

El concepto de d-separación se corresponde con el de independencia condicional, de manera que dos variables (o conjuntos de variables) X e Y serán condicionalmente independientes dada una tercera variable (o conjunto de variables) Z si y sólo si Z d-separa a X e Y.

2.1. Construcción de las redes a partir de los datos

Existen diversas técnicas para construir redes bayesianas a partir de una base de datos. En este trabajo hemos empleado el algoritmo K2 [1], que está basado en la optimización de una medida. Esa medida se usa para explorar, mediante un algoritmo de ascensión de colinas, el espacio de búsqueda formado por todas las redes que contienen las variables de la base de datos. Se parte de una red inicial y ésta se va modificando (añadiendo arcos, borrándolos o cambiándolos de dirección) obteniendo una nueva red con mejor medida. En concreto, la medida K2 [1] para una red G y una base de datos D es la siguiente:

$$\begin{aligned} f(G:D) &= \log P(G) \\ &+ \sum_{i=1}^{n}\left[\sum_{k=1}^{s_i}\left[\log\frac{\Gamma(\eta_{ik})}{\Gamma(N_{ik}+\eta_{ik})}\right.\right. \\ &+ \left.\left.\sum_{j=1}^{r_i}\log\frac{\Gamma(N_{ijk}+\eta_{ijk})}{\Gamma(\eta_{ijk})}\right]\right] , \end{aligned}$$

donde N_{ijk} es la frecuencia de las configuraciones encontradas en la base de datos D de las variables x_i, donde n es el número de variables, tomando su j-ésimo valor y sus padres en G tomando su k-ésima configuración, donde s_i es el número de configuraciones posibles del conjunto de padres y r_i es el número de valores que puede tomar la variable x_i. Además, $N_{ik} = \sum_{j=1}^{r_i} N_{ijk}$ y Γ es la función Gamma.

El algoritmo K2 opera sobre variables discretas, por lo que las continuas han de ser discretizadas previamente. En este trabajo hemos usado el método k-means de clustering para realizar la discretización, con $k = 5$.

3. Descripción de los datos

Los datos objeto de estudio han sido los referentes a rendimiento académico de los estudiantes de la Universidad de Almería duran-

te el curso 2003/2004. Estos datos están compuestos por los valores de las variables que a continuación se relacionan para todas las asignaturas impartidas en dicho curso, que suponían un total de 1346.

- **TC:** Porcentaje de alumnos a tiempo completo en la titulación donde se imparte la asignatura, es decir, matriculados en 50 créditos o más de esa titulación.
- **MediaAcceso:** Nota media de acceso de los alumnos matriculados en la asignatura.
- **P80Acceso:** Percentil 80 de la nota de acceso de los alumnos de la asignatura.
- **Alum_GT:** Número de alumnos por grupo de teoría.
- **Alum_GP:** Número de alumnos por grupo de prácticas.
- **Titulación:** Código de la titulación en la que se oferta la asignatura.
- **Repetidores:** Porcentaje de repetidores en la asignatura.
- **ConvUtilizadas:** Media de convocatorias utilizadas para aprobar la asignatura.
- **IndiceUtilConv:** 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.
- **TasaAproba:** Porcentaje de alumnos con calificación .ªprobado".
- **TasaNotable:** Porcentaje de alumnos con calificación "Notable".
- **TasaSobre:** Porcentaje de alumnos con calificación "Sobresaliente".
- **CalifRela:** Calificación relativa de la asignatura.
- **TasaÉxito:** Número de alumnos aptos dividido por número de alumnos presentados.
- **TasaRendimiento:** Número de alumnos aptos dividido por número de alumnos matriculados.
- **NumProf:** Número de profesores distintos en la asignatura.
- **EvaluaProf:** Media de las notas obtenidas en las encuestas de los profesores participantes en la asignatura.
- **Créditos_Doctor:** Porcentaje de créditos impartidos por doctores en la asignatura.
- **ContratoProf:** Porcentaje de créditos impartidos por profesores funcionarios en la asignatura.

De entre todas las variables, es especialmente interesante tratar de descubrir los aspectos contenidos en los datos que determinan la tasa de éxito y la tasa de rendimiento, dos de los indicadores más importantes desde el punto de vista de la gestión universitaria.

Partiendo de estos datos, hemos estudiado cuatro escenarios distintos, y en cada uno de ellos hemos construido una red bayesiana a partir de los correspondientes datos. Los escenarios son los siguientes:

- **Escenario 1:** Asignaturas troncales y obligatorias. A la red correspondiente la llamaremos Red 1.
- **Escenario 2:** Asignaturas optativas. Denotaremos por Red 2 a la red que representa esta situación.
- **Escenario 3:** Asignaturas troncales, obligatorias y optativas, añadiendo además una variable nueva, CAR, que indica el carácter de la asignatura. La correspondiente red es Red 3.
- **Escenario 4:** Igual que el Escenario 3, pero sustituyendo los posibles valores de la variable titulación por el tipo de estudio (ciclo corto, ciclo largo o sólo segundo

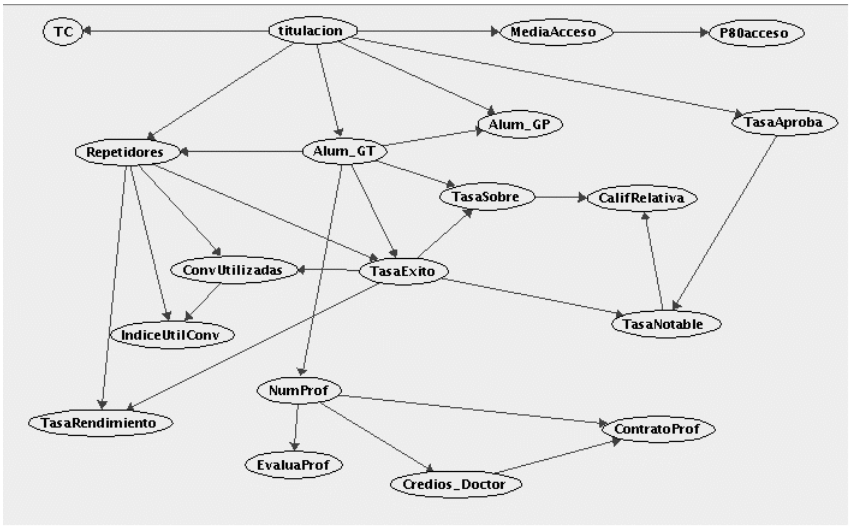

Figura 2: Red obtenida para asignaturas troncales y obligatorias.

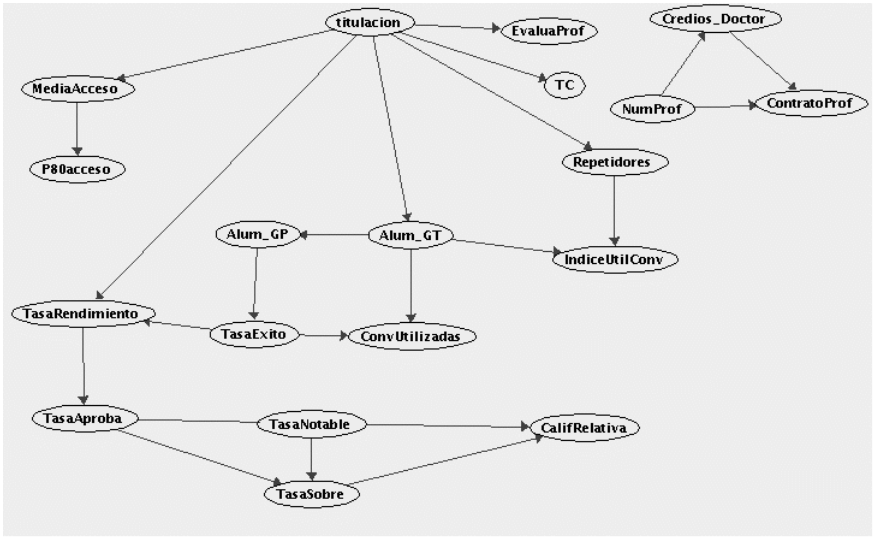

Figura 3: Red obtenida para asignaturas optativas.

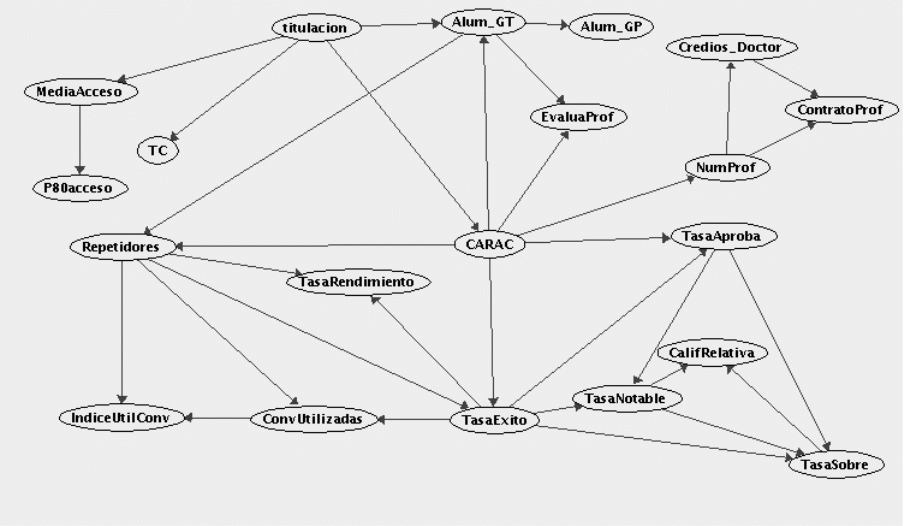

Figura 4: Red obtenida para los tres tipos de asignaturas.

ciclo). En este caso, la red será denotada como Red 4.

Las redes obtenidas pueden verse en las figuras 2 a 5.

4. Análisis de los resultados proporcionados por las redes

En primer lugar analizaremos los escenarios 1 y 2 a través de las diferencias entre las dos primeras redes, la de troncales y obligatorias frente a la de optativas. En ésta última, se observa que la calificación obtenida por el profesor en la evaluación, depende exclusivamente de la titulación cuando el valor de ésta es conocido, mientras que en el caso de la Red 1, es el número de profesores que imparten la asignatura lo que determina la evaluación obtenida.

Cabe destacar que en el caso de las optativas (Red 2), la información relativa al tipo de profesorado y su número no es relevante para ninguno de los demás indicadores, y que en el caso de la Red 1, éstos parámetros sólo influyen de forma indirecta sobre los indicadores fundamentales, a través de la variable Alum_GT (número de alumnos por grupo de teoría), de manera que si se conoce el valor de ésta variable, la información acerca del profesor no determina en absoluto la tasa de éxito.

En ambas redes, el porcentaje de repetidores está fuertemente relacionado con la titulación. Realizando propagación de probabilidades [4], puede observarse, sin embargo, que el porcentaje de repetidores es significativamente más elevado en las asignaturas troncales y obligatorias. Esto era de esperar, ya que en el caso de las optativas, es frecuente que alumnos que suspenden una asignatura, al curso siguiente se matriculen en otra distinta en lugar de volver a intentar aprobar la misma asignatura.

Se observan también diferencias entre ambas redes en cuanto al índice de utilización de convocatorias. En la Red 1, se observa que este índice depende del porcentaje de repetidores y del número de convocatorias utilizadas, mientras que en el caso de las optativas, en lugar del número de convocatorias utilizadas, depende del número de alumnos por grupo de teoría, ya que en el caso de las optativas, la distribución del número de convocatorias utilizadas es muy homogénea.

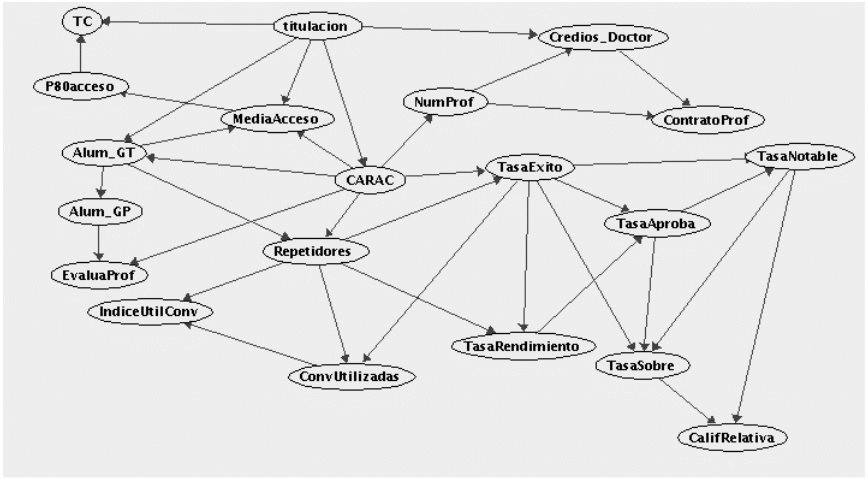

Figura 5: Red con todas las asignaturas y tipo de estudio.

Otra diferencia destacable es que la tasa de aprobados depende depende directamente de la titulación en la que se ofertan las asignaturas troncales y obligatorias, mientras que para las optativas dependen directamente de la tasa de rendimiento, y sólo indirectamente de la titulación.

En cuanto a los indicadores principales, la tasa de éxito y la de rendimiento, ambos están directamente relacionados tanto en la Red 1 como en la Red 2, lo que indica la existencia de una correlación significativa entre ambas.

En el caso de las optativas, la tasa de éxito está relacionada con el número de alumnos por grupo de prácticas, mientras que para las troncales y obligatorias es con el número de alumnos por grupo de teoría con el que tiene una relación directa. Esto puede deberse a que en las optativas normalmente sólo hay un grupo de teoría, mientras que en troncales y obligatorias, esta cifra varía entre 1 y 5 dependiendo de la titulación y el curso de la asignatura.

En la Red 1, influyen directamente sobre la tasa de éxito el porcentaje de repetidores, la media de convocatorias utilizadas para aprobar y, a través del número de alumnos por grupo de teoría, del número de profesores y de la titulación. Es obvia la relación encontrada entre el porcentaje de repetidores y la tasa de éxito de una asignatura: a menor tasa de éxito, mayor número de repetidores.

Otras relaciones destacables son las siguientes:

- En la Red 1, la titulación influye directamente sobre la tasa de aprobados y, a través de ella, en la tasa de notables y en la calificación relativa de la asignatura. En el caso de las optativas, esta relación pasa por la variable tasa de rendimiento.
- En ambas redes, el nivel de acceso de los alumnos, representado por las variables media de acceso y percentil 80 de acceso, no influye en las tasas de éxito/rendimiento una vez conocida la titulación a la que pertenece la asignatura.
- Independientemente del tipo de asignatura, la titulación está directamente relacionada con el porcentaje de repetidores.
- También en ambos casos, el número de alumnos por grupo se muestra como un

factor decisivo en la tasa de éxito y, a través de ésta, en la tasa de rendimiento.

- El porcentaje de alumnos que se dedican a tiempo completo a la titulación donde se imparte la asignatura, depende directamente sólo de la titulación, por lo que si ésta es conocida, no tiene ninguna influencia sobre las tasas de éxito y rendimiento.

En cuanto a los escenarios 3 y 4, estos son bastante similares. Cabe destacar que la introducción del carácter de la asignatura aporta información sobre los resultados de las evaluaciones del profesorado, aspecto que no podía ser detectado en los dos primeros escenarios.

5. Conclusiones

En este trabajo hemos presentado un análisis de algunos indicadores relacionados con el rendimiento, tomando como base únicamente la información aportada por los datos. Pensamos que este estudio pone de manifiesto la validez de las redes bayesianas para abordar situaciones que pueden ser de especial interés en gestión universitaria. Hay que tener en cuenta que este tipo de modelos proporcionan una salida fácilmente interpretable por un usuario sin formación estadística, ya que de forma visual permite detectar las variables que están relacionadas entre sí en base al concepto de d-separación.

Pretendemos ampliar este estudio con un análisis detallado de las relaciones detectadas entre indicadores, mediante propagación de probabilidad [4] y extracción de perfiles [2], con objeto de cuantificar las dependencias entre las variables objeto de estudio.

Referencias

[1] G.F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[2] L.M. de Campos, J.A. Gámez, and S. Moral. Partial abductive inference in Bayesian networks by using probability trees. In *Proceedings of the 5th International Conference on Enterprise Information Systems (ICEIS'03)*, pages 83–91, Angers, 2003.

[3] Juan Hernández. *La Universidad española en cifras*. CRUE, 2004.

[4] Finn V. Jensen. *Bayesian networks and decision graphs*. Springer, 2001.

[5] M. Morales and A. Salmerón. Análisis del alumnado de la Universidad de Almería mediante redes bayesianas. In *Actas del 27 Congreso Nacional de Estadística e I.O.*, pages 3413–3436, 2003.

[6] J. Pearl. *Probabilistic reasoning in intelligent systems*. Morgan-Kaufmann (San Mateo), 1988.