*Revista de Investigación Educativa, 2003, Vol. 21, n.º 1, págs. 93-111*

# EL USO DE LAS TÉCNICAS DE SEGMENTACIÓN EN LA EVALUACIÓN DEL RENDIMIENTO EN LENGUAS. UN ESTUDIO EN LA COMUNIDAD AUTÓNOMA VASCA[1]

*L. Lizasoain, L. Joaristi, C. Santiago, J.F. Lukas, N. Moyano, M. Sedano, B. Munárriz*[2]

## RESUMEN

*Este artículo es una ilustración del empleo de las técnicas estadísticas de segmentación en el análisis de los datos de una investigación evaluativa. Se ha aplicado el método CART a un conjunto de datos sobre rendimiento académico en lengua vasca y española pertenecientes a una muestra de estudiantes de enseñanza secundaria de la Comunidad Autónoma Vasca. Los objetivos eran diseñar y depurar un modelo predictivo del rendimiento en estas materias, así como valorar las posibilidades que el uso de este tipo de técnicas ofrecen tanto en la fase del análisis de los datos de las investigaciones evaluativas como en la de la comunicación de los resultados. Los resultados obtenidos mediante segmentación han sido triangulados usando la regresión múltiple, el análisis de componentes principales y el análisis de correspondencias. Los resultados son básicamente coincidentes, pero la segmentación ofrece la ventaja de poder operar simultáneamente tanto con variables cuantitativas como cualitativas. Además los resultados gráficos que ofrece (árboles de decisión) son de muy sencilla interpretación.*

***Palabras clave**: evaluación de centros, educación secundaria, rendimiento académico, lengua vasca, lengua española, técnicas de segmentación, árboles de decisión.*

---

1 Esta investigación ha sido subvencionada en parte por el Departamento de Educación del Gobierno Vasco-Eusko Jaurlaritza, la Universidad del País Vasco-Euskal Herriko Unibertsitatea (proyecto 1/UPV/EHU 00218.230-HA-8114/2000) y el Instituto de Evaluación y Asesoramiento Educativo (IDEA).

2 Universidad del País Vasco-Euskal Herriko Unibertsitatea. Departamento de Métodos de Investigación y Diagnóstico en Educación. Facultad de Filosofía y Ciencias de la Educación. Avda. de Tolosa 70. 20018 San Sebastián. E-mail: plplihel@sf.ehu.es

## ABSTRACT

*This paper is an illustration of the use of segmentation statistical techniques in the data analysis of an evaluative research. The CART method has been applied to a set of data on academic achievement in Basque and Spanish language. The data set was obtained from a sample of Secondary Education students in the Basque Autonomous Community. The aims were to devise and purge a predictive model of the achievement in these areas, and to assess the possibilities that the use of this kind of technique offers both in the data analysis phase of evaluative research and in that of dissemination of results. The results obtained through segmentation have been triangulated by means of multiple regression, factor analysis and multiple correspondence analysis. The results basically agree, but segmentation has the advantage that one can operate simultaneously with both quantitative and qualitative variables. Furthermore, the graphical results offered (decision trees) are very easy to interpret.*

***Keywords**: school assessment, secondary education, academic achievement, basque language, spanish language, segmentation techniques, decision trees.*

## INTRODUCCIÓN Y OBJETIVOS

Desde el curso 1999-2000 se lleva realizando en la Comunidad Autónoma Vasca (CAV) una evaluación externa de centros de educación secundaria promovida y financiada por el Gobierno Vasco, desarrollada por profesores e investigadores del departamento de Métodos de Investigación y Diagnóstico en Educación de la Universidad del País Vasco-Euskal Herriko Unibertsitatea y con la colaboración del Instituto de Evaluación y Asesoramiento Educativo (IDEA).

Este proyecto evaluativo se inserta en el contexto general de la Red de Evaluación de centros (REDES) que desde 1997 lleva a cabo evaluaciones externas en cerca de 150 centros de educación secundaria.

Información más detallada sobre las cuestiones relativas a las características de dicho proyecto de evaluación, sus niveles y dimensiones, junto con diversos resultados obtenidos; pueden encontrarse en Equipo REDES (1999 y 2000), y más recientemente, en Marchesi, A. y Martín, E. (comp.) (2002).

En nuestro caso, el objetivo de la evaluación que estamos desarrollando es que cada uno de los centros participantes obtenga información objetiva, fiable y válida acerca de su funcionamiento y de sus resultados de forma que le posibilite tomar las decisiones que consideren oportunas de cara a mejorar el propio funcionamiento y rendimiento de su alumnado. Es una evaluación externa que no pretende suplantar la evaluación o evaluaciones que el profesorado lleva a cabo en su centro correspondiente. Sin embargo, esta evaluación puede ser utilizada como complementaria a la evaluación interna de cada centro, en el sentido de ofrecer un conjunto de datos objetivos, cuya interpretación y valoración deben ser realizadas por el personal del propio centro tomando en consideración sus propios criterios.

Desde esta perspectiva, como **primer objetivo** de este trabajo nos planteamos el exponer la metodología y primeros resultados obtenidos en la fase preliminar de esta

evaluación de centros de la ESO. Más en concreto, el estudio realizado pretende ***diseñar y depurar un modelo predictivo del rendimiento en las materias de lengua española y lengua vasca.***

Adicionalmente y como **segundo objetivo**, pretendemos ***mostrar las posibilidades que las técnicas de segmentación*** ofrecen en este ámbito y para los fines analíticos planteados.

Este conjunto de técnicas, también conocidas como árboles de decisión, no son muy empleadas en nuestro campo disciplinar.

Las búsquedas documentales realizadas arrojan un resultado exiguo pues sólo hemos encontrado los trabajos de Everett y otros (1997); Godley, Fiedler y Funk (1998); y Forthofer y Bryant (2000). En ellos se elaboran, analizan y validan modelos predictivos sobre los factores académicos en escuelas rurales, sobre el grado de satisfacción con los servicios de salud mental o para evaluar el ajuste de estrategias de cambio conductual.

En nuestro entorno más cercano, Repetto y otros (1994) emplearon las técnicas de segmentación en la evaluación de un programa de orientación metacognitiva de la comprensión lectora.

Se trata de un conjunto de técnicas que permiten definir y validar modelos de forma que se pueda determinar qué variables (predictoras) inciden o explican la variabilidad de una variable dependiente.

Son, por tanto, técnicas explicativas de la familia de la regresión o el análisis discriminante pero tienen la ventaja de que tanto la variable criterio como las predictoras pueden ser de cualquier tipo (tanto cuantitativas como cualitativas) lo que en nuestro contexto es especialmente importante.

La mayoría de los autores coinciden en que con estas técnicas es posible abordar problemas y cuestiones como la propia segmentación de poblaciones, la validación de modelos predictivos, la reducción de la dimensionalidad o la identificación de la interacción. Se trata, como vemos, de problemas que suelen plantearse con frecuencia en la fase del análisis de los datos de los estudios evaluativos.

Pero además de sus capacidades analíticas y de poder operar con cualquier tipo de variables, otra de las principales ventajas que estas técnicas aportan estriba en que sus resultados se presentan de forma gráfica (árboles de decisión) siendo de muy sencilla interpretación.

Y esto es de crucial importancia cuando hay que presentar los resultados de un estudio evaluativo a públicos no expertos. Autores como Patton (1997) o Henry (1993, 1998) afirman que los implicados no expertos en estadística o análisis de datos pueden comprender e interpretar datos y resultados cuando son presentados en forma clara y legible usando gráficos y tablas estadísticas. En otro lugar (Lizasoain y Joaristi, 2000) hemos afirmado que en evaluación de programas, después de una *primera vuelta* de análisis e interpretación, es necesario *re-analizar* los resultados para hacerlos comprensibles en la presentación.

Evidentemente, este conjunto de técnicas no constituye ninguna panacea pues también tiene limitaciones. Como pretendemos mostrar en este artículo, además de su uso en la interpretación y comunicación de resultados, lo más apropiado es emplear estas

técnicas para diseñar y depurar modelos que puedan ser luego analizados mediante técnicas inferenciales más potentes.

## CARACTERÍSTICAS Y DISEÑO DEL PROYECTO DE EVALUACIÓN

Esta evaluación se ha realizado en 55 centros de Enseñanza Secundaria Obligatoria (ESO) de la Comunidad Autónoma Vasca. En primer lugar, hemos de señalar que la participación de los centros es totalmente voluntaria, por lo que no podemos hablar de una muestra representativa de la población escolar de secundaria de la CAV dado que no ha habido aleatoriedad en la selección. De todas formas, teniendo en cuenta la amplitud de dicha muestra y el esfuerzo realizado para tratar de controlar su distribución en función de variables como el contexto sociocultural, la titularidad, etc., consideramos que los datos obtenidos pueden ser útiles para proporcionarnos una idea de lo que puede estar ocurriendo en la Enseñanza Secundaria Obligatoria en la CAV (Santiago y Lukas, 2000).

En cuanto a la titularidad de los centros, son 17 centros públicos (30%) y 38 centros privados concertados (70%). Pero si estos datos los examinamos considerando la proporción por aulas obtenemos que el total se distribuye a medias entre un 50% de pertenecientes a centros públicos y otro tanto a centros privados. En el curso 1999-2000 han participando alrededor de 8000 estudiantes de 1º, 2º y 4º de ESO. De cualquier forma, está previsto que este plan de evaluación abarque 4 cursos académicos consecutivos.

Con respecto a las variables que se consideran objeto de estudio, se agrupan en tres bloques:

- Un primer conjunto está formado por el rendimiento escolar en diferentes disciplinas: Lengua y Literatura Vasca, Lengua y Literatura Española, Matemáticas, Ciencias Sociales y Ciencias Naturales.
- En segundo lugar, se han incorporado aspectos como las actitudes, el grado o nivel de satisfacción, las estrategias de aprendizaje y las habilidades metacognitivas.
- Un tercer bloque está integrado por las cuestiones relativas al contexto económico, social y cultural en que se insertan los centros y los estudiantes y sus familias.

La obtención de los datos de rendimiento se realiza mediante la aplicación de pruebas curriculares. Se trata de pruebas objetivas que tratan de mantener un equilibrio entre los contenidos conceptuales y los procedimentales.

Éstas son preparadas en una primera fase en el Instituto IDEA (Instituto de Evaluación y Asesoramiento Educativo) por expertos (profesorado de secundaria y personal experto en construcción de pruebas) que pilotan dichas pruebas y realizan los correspondientes análisis de ítems y de la prueba, siguiendo modelos de Teoría Clásica de los Tests y de la Teoría de la Respuesta al Item. En nuestro caso, en la CAV, las pruebas diseñadas son revisadas por una red constituida por profesorado de Primaria y Secundaria que analiza la adecuación de los ítems a las líneas curriculares de cada área, por profesorado universitario especializado en Medición y Evaluación y por el –hasta hace unos meses denominado– Instituto de Desarrollo Curricular (IDC) dependiente del

Gobierno Vasco. Con la información aportada por estos tres grupos de expertos, se realiza un informe que es enviado a IDEA para la confección definitiva de cada prueba. Posteriormente son traducidas a la lengua vasca y enviadas para ser impresas. En todo este proceso se han seguido las directrices señaladas por la *International Test Comission* (Hambleton, 1996) para la traducción y adaptación de las pruebas.

Siguiendo la clasificación de Jornet y Suárez (1996), podemos definir las pruebas utilizadas como *Pruebas Estandarizadas de Indicadores de Resultados*. Este tipo de pruebas pretenden traducir los niveles de competencia que en las distintas disciplinas y materias una institución, centro o sistema asume como objetivos a cumplir en el proceso educativo.

En la parte quinta de la obra citada de Marchesi y Martin (2002), se pueden encontrar descripciones más detalladas de los procedimientos de obtención de datos en las áreas curriculares de Matemáticas, Lengua, Ciencias Sociales, Ciencias de la Naturaleza y Tecnología. (Cada uno de los 5 capítulos, del 9º al 13º, aborda una materia).

Además de los resultados en las materias citadas, los estudiantes fueron evaluados también en lo tocante a las habilidades metacognitivas y a las estrategias de aprendizaje mediante la aplicación de cuestionarios.

En el primer caso, el cuestionario pretende evaluar cuatro tipo de procesos: la metacomprensión, la verificación de resultados, la conciencia de las estrategias utilizadas y la conciencia del propio conocimiento. En lo relativo a las estrategias se distinguen cinco factores: el dominio estratégico, la elaboración de resúmenes y esquemas, la reflexión sobre el propio trabajo, el establecimiento de relaciones y el empleo de la memorización.

Para finalizar con la medición de las variables del segundo bloque, en éste se incluye la aplicación de cuestionarios de opinión para medir el grado de satisfacción con el centro tanto de los alumnos como de los padres. Los cuestionarios aplicados, son en su mayoría, escalas de tipo Likert.

En el tercer bloque se incluyen las variables relativas al contexto sociocultural. Mediante la aplicación de un cuestionario se obtiene un índice de nivel familiar a través de indicadores como el nivel de estudios y la situación profesional de los padres, el número de coches en el hogar, el número de libros, la presencia o no de ordenador, etc. Posteriormente, tras ponderar los indicadores, se obtiene el índice de cada alumno y a continuación el índice promedio del centro.

Pero además de esto, en este tercer apartado se incluye una variable a la que debemos hacer mención expresa, pues se trata de una variable muy importante y específica del sistema educativo vasco: el modelo lingüístico. En función de cuál sea la lengua vehicular del proceso de enseñanza-aprendizaje, existen tres modelos: A, B y D.

En el modelo «A» la enseñanza se desarrolla en castellano y la lengua vasca se enseña como una asignatura, mientras que en el modelo «D» la situación es la inversa: todas las materias se imparten en euskara y el castellano es una asignatura. El modelo «B» es el intermedio, y en el mismo unas materias se enseñan en vasco y otras en español.

Como más adelante veremos, además de ser una especificidad importante de nuestro sistema educativo, esta variable juega un papel relevante en el modelo que proponemos.

Tomando en consideración todos estos antecedentes, el estudio que nos planteamos se plantea como objetivo básico el diseñar un modelo predictivo que estudie la relación entre el rendimiento escolar en lengua española y lengua vasca y el resto de las variables consideradas en el estudio.

En concreto, los datos que vamos a emplear provienen de 2143 estudiantes de primer curso de ESO de quienes se han incorporado las siguientes variables:

- Modelo lingüístico (MODELO) (A, B y D)
- Sexo (SEXO)
- De lo relativo a las actitudes hemos considerado los 4 aspectos siguientes:
  — Tolerancia (TOLERFIN)
  — Ecología (ECOLOFIN)
  — Transversalidad (TRANSFIN)
  — Salud (SALUDFIN)
- Contexto del centro (CONTEXTO) con cuatro valores (bajo, medio bajo, medio alto, y alto).
- Puntuación directa del contexto sociocultural del alumno (PUNTUACI)
- Puntuación total de habilidades metacognitivas (HABILIDA)
- Puntuación en estrategias-dominio (DOMIFIN)
- Puntuación en estrategias-esquemas (ESQUEFIN)
- Puntuación en estrategias-reflexión (REFLEFIN)
- Puntuación en estrategias-relaciones (RELAFIN)
- Puntuación en estrategias-memorístico (MEMOFIN)
- Puntuación de la pueba de castellano (LNOTAFIN)
- Puntuación de la prueba de euskara (KNOTAFIN)
- Tipo de centro (TIPOCEN) (público, privado)

Pero además, y desde la perspectiva metodológica que hemos apuntado como segundo objetivo básico, nos interesa valorar la aplicabilidad de las técnicas de segmentación a este tipo de problemas. Y para poder cumplir tal objetivo y triangular los resultados, es necesario comparar este conjunto de técnicas con las, digamos, habituales o *clásicas* en este tipo de estudios: las técnicas factoriales y las de regresión múltiple.

## APLICACIÓN DE LAS TÉCNICAS *CLÁSICAS*

Como ha quedado dicho, nuestro objetivo es proponer un modelo que estudie la relación entre el rendimiento en lengua castellana y en lengua vasca con el resto de las variables consideradas en la investigación.

### Análisis de la dimensionalidad

En una primera maniobra de aproximación al problema, antes de diseñar un modelo predictivo conviene explorar la estructura dimensional. Para ello realizamos un análisis de componentes principales tanto con las variables que consideramos como

dependientes como con las predictoras, aunque con la lógica limitación de poder incluir de éstas sólo las cuantitativas.

La figura 1 muestra el gráfico de componentes en el espacio rotado (Varimax). Como puede observarse, 4 de los factores de las distintas *estrategias de aprendizaje* se sitúan en torno al primer componente (39,9%), mientras que el *contexto del estudiante*, el *total de las habilidades metacognitivas* y las *estrategias memorísticas* se correlacionan con el segundo componente (19%), estando ésta última variable en clara oposición a las anteriores.

Por su parte, las puntuaciones en *lengua española* y *lengua vasca*, tratadas como variables suplementarias, se sitúan próximas a este segundo componente.

En resumen, entre las estrategias de aprendizaje relacionadas con el dominio, la reflexión, el uso de esquemas y las relaciones hay ortogonalidad respecto a la puntuación total en las habilidades metacognitivas, el contexto del alumno y el empleo de la memoria como estrategia de aprendizaje. Las variables dependientes, rendi-

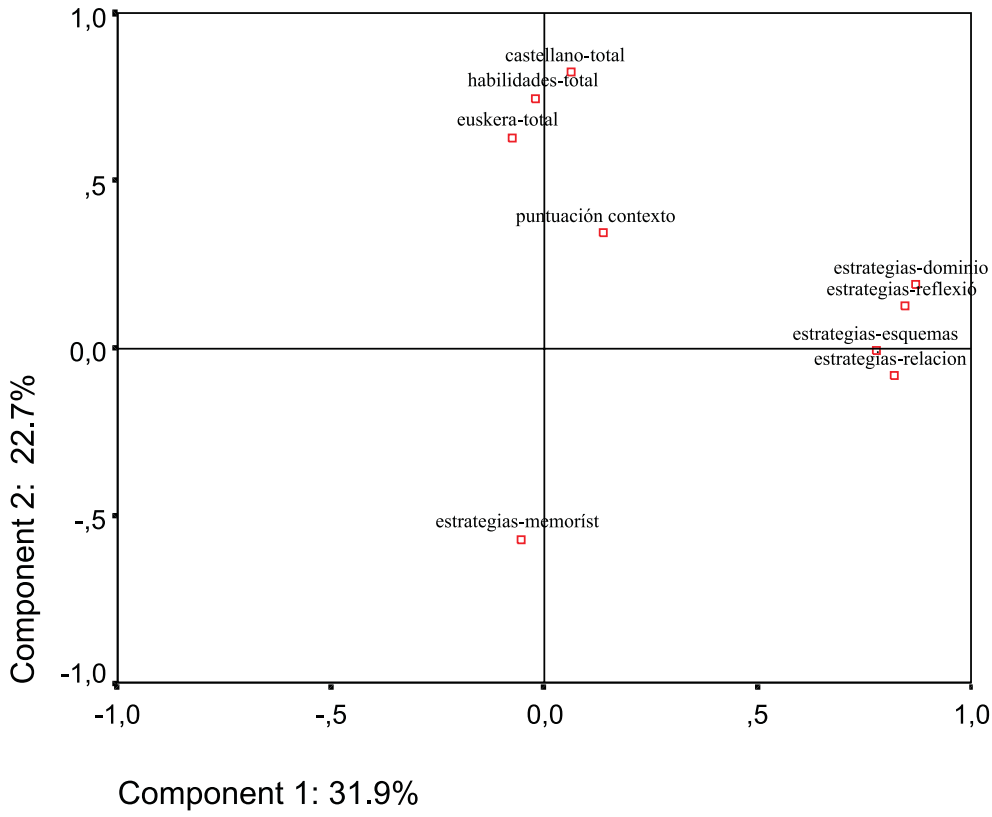


Figura 1
*Plano factorial resultante de la aplicación del Análisis de Componentes Principales.*

miento en lengua española y lengua vasca, ésta en menor medida, están asociadas a éstas últimas.

Como primera conclusión diríamos que entre las variables independientes cuantitativas serían *Habilidades, Estrategias memorísticas y el Contexto del alumno* las mejores predictoras. Todo ello con la limitación evidente que antes apuntábamos de no poder contar con las variables predictoras cualitativas.

Para poder solventar este problema una estrategia plausible es recurrir al análisis de correspondencias múltiples previa categorización de las variables cuantitativas, en este caso en 4 niveles equiponderados.

Una vez realizado el análisis con todas las variables como activas, se obtiene el primer plano factorial (figura 2). En el mismo, tras realizar la corrección de Benzécri, las nuevas tasas de inercia de los 2 primeros factores son 59,78% para el primer factor y 19,21% para el segundo.

Aunque en esta figura aparecen las etiquetas de las modalidades, la interpretación es como sigue: las variables relativas a las estrategias de dominio, reflexión y esquemas están asociadas al primer factor, mientras que el rendimiento en castellano y en euskara, así como el total de las habilidades metacognitivas se asocian al segundo. Esta estructura factorial coincide plenamente con la que acabamos de ver resulante del análisis de componentes principales.

En lo relativo a las variables cualitativas modelo lingüístico y contexto sociocultural, lo más relevante es que los modelos A y B se encuentran en el semiplano de las pun-

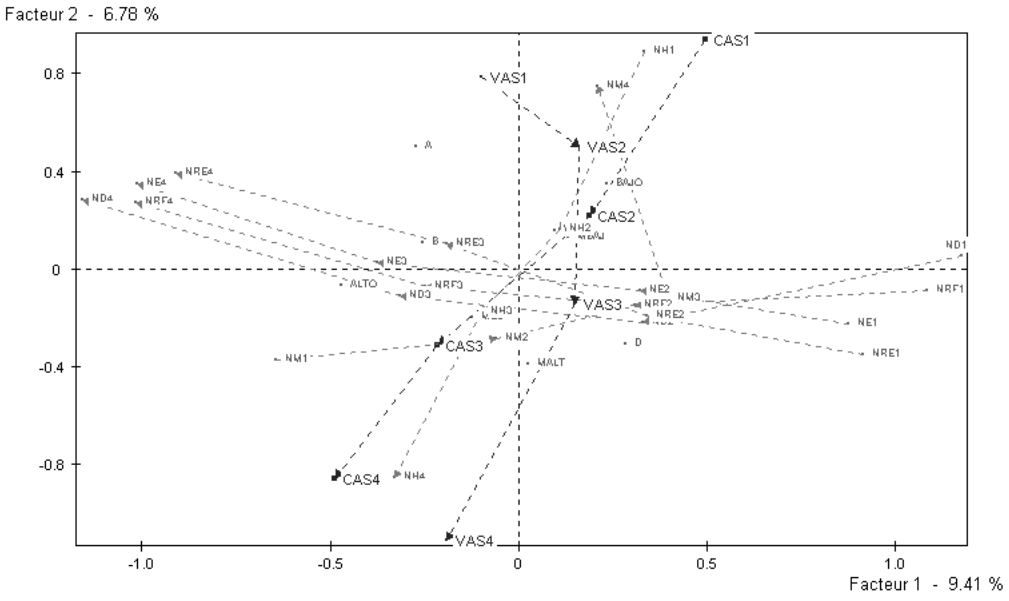


Figura 2

*Plano factorial resultante de la aplicación del Análisis de Correspondencias Múltiples.*

tuaciones bajas en rendimiento en euskara y castellano. Igualmente, cerca de estas puntuaciones se sitúan las modalidades baja y medio-baja de la variable contexto.

En conclusión, los resultados obtenidos apuntan a una estructura bidimensional:

- un primer factor formado por las estrategias de aprendizaje de dominio, reflexión, esquemas y relaciones que es ortogonal al
- segundo factor integrado por el rendimiento en ambas lenguas, el total de las habilidades metacognitivas, el contexto, los modelos y el empleo de estrategias memorísticas. Aquí los valores bajos del contexto, los modelos A y B y las estrategias memorísticas se sitúan en el semiplano izquierdo.

## Regesión lineal múltiple

Como ha quedado dicho, nuestro objetivo es proponer un modelo que estudie la relación entre el rendimiento en lengua castellana y en lengua vasca con el resto de las variables consideradas en la investigación.

Desde una perspectiva —digamos *clásica*— este problema se aborda mediante la regresión múltiple. En este caso la variable dependiente es el rendimiento en cada una de las lenguas y las predictoras el resto de las variables.

En este punto se plantea el problema de las variables predictoras cualitativas (el contexto, el modelo lingüístico y el género), que se supera mediante la generación de las variables ficticias (*dummy*) correspondientes.

En ambos casos hemos procedido a efectuar la regresión múltiple por pasos una vez verificado –para nuestra sorpresa– que se cumplen estrictamente todas las condiciones de aplicación. La tabla 1 muestra los resultados del último paso para el rendimiento en lengua castellana.

Como vemos, en esta última etapa han sido incluidas 9 variables, de las que por su peso beta destacamos la puntuación total en habilidades metacognitivas, las estrategias memorísticas (con una relación inversa) y en tercer lugar las estrategias de dominio.

En la tabla 2 se muestran los resultados de la regresión múltiple sobre el rendimiento en lengua vasca.

Aquí nos encontramos con unos resultados distintos. Junto con la puntuación total en habilidades metacognitivas y las estrategias memorísticas, aparecen dos nuevas variables (Modelo A y Modelo B) ambas relacionadas inversamente con el rendimiento. Hay que hacer notar que Modelo A es la primera variable en ser incluida en el modelo, con una beta de -0,653. Por su parte, Modelo B aparece en tercer lugar con una beta de -0,211. Esto, de acuerdo a la codificación binaria realizada previamente para generar las variables ficticias, se interpreta como que los estudiantes del modelo A obtienen un rendimiento en lengua vasca inferior a los del modelo B y éstos, a su vez, inferior a los del modelo D.

De los análisis de regresión cabe resumir las siguientes conclusiones:

- La puntuación total en habilidades metacognitivas (en sentido positivo), las estrategias memorísticas (en sentido negativo) y las estrategias de dominio son las mejores predictoras del rendimiento en ambas lenguas.

- Sin embargo, para el caso de la lengua vasca es el modelo lingüístico la variable que adopta un lugar destacado, oponiendo el modelo A al modelo D.

Por tanto, desde el punto de vista de las habilidades metacognitivas y las estrategias de aprendizaje, son la puntuación total en las primeras y las de tipo memorístico en las segundas, las variables que más explican la variación del rendimiento en las lenguas objeto de estudio.

Adicionalmente, el modelo propuesto para el rendimiento en lengua vasca tiene mayor capacidad explicativa ($R^2$ = 0,481 frente a 0,36). Este incremento es debido a la incorporación de la variable modelo lingüístico.

## APLICACIÓN DE LAS TÉCNICAS DE SEGMENTACIÓN

Con estos antecedentes, y vistas las dificultades que ocasiona la coexistencia de variables cualitativas y cuantitativas, examinemos ahora los resultados de la aplicación de las técnicas de segmentación con el objeto de ver si el uso de las mismas puede resultar de ayuda en este tipo de contextos. De nuevo aquí, vamos a operar con el mismo conjunto de variables en las que el rendimiento global de los sujetos en las materias de Lengua Española y Lengua Vasca van a actuar como variables dependientes. El análisis de segmentación se ha realizado mediante el programa *Answer Tree* que distribuye SPSS (1999) como módulo adicional. Y para llevar a cabo dicho análisis es preciso especificar las siguientes cuestiones:

- **Elección del método o algoritmo de segmentación.**

Cuatro son los principales algoritmos que se pueden emplear en este tipo de análisis, y, en este caso, dado el tipo de variables con las que se cuenta, los más indicados son el CHAID exhaustivo (Kass, 1980) (Bigs, de Ville y Suen, 1991) y el CART (Breiman y otros, 1984).

Hemos optado por este segundo porque produce árboles binarios que resultan, en principio, más fáciles de interpretar. Además, después de comparar ambos con estos datos, este algoritmo produce mejores soluciones en el sentido de que consigue explicar más varianza. No es que haya una gran diferencia pero ésta aparece siempre a favor del método CART.

- **Definición de las variables.**

En cualquier caso, y sea cual sea el algoritmo elegido, debe procederse a la definición de las variables que se incorporan al modelo especificando cuál es la variable dependiente y cuáles van a actuar como predictores sin que se plantee ningún tipo de restricción en función de su nivel de medida.

En consecuencia, las ***variables predictoras*** son la puntuación total en *Habilidades*, las diferentes *Estrategias (dominio, esquemas, reflexión, relaciones, memoria), el Género* de los alumnos, el *Modelo lingüístico* en el que cursan los estudios y el *Contexto socio-económico* tanto del alumno como del centro (puntuación contexto alumno y contexto).

Como ***variables dependientes***, al igual que antes vamos a comparar los resultados en *Lengua Española y Lengua Vasca*.

- **Especificación del procedimiento de validación.**

Como los distintos autores señalan, es recomendable validar el árbol con objeto de incrementar su precisión (lo que en algunas fuentes se denomina *validez predictiva*). Además los árboles no validados tienden a subestimar el riesgo de clasificación o predicción errónea.

En este caso, dado que el tamaño de la muestra (2143 casos) nos lo permite, vamos a partir la muestra en dos submuestras con las proporciones habitualmente recomendadas: la muestra de aprendizaje con el 67% de los casos (n = 1441) y la de prueba o validación con el 33% restante (n = 702).

### Generación del árbol

En el momento en que se finaliza la especificación de parámetros, el programa genera y muestra el árbol mínimo, el *nodo-raíz*. A partir de aquí, son posibles tres procedimientos de generación del árbol:

- Generarlo de una sola vez (*de golpe*) hasta que se alcance uno de los criterios de parada.
- Generarlo haciéndolo crecer nivel a nivel.
- Generarlo hasta el final y luego *podarlo*. La *poda* consiste en una optimización del modelo propuesto mediante la eliminación de ramas y nodos que incrementan la complejidad del modelo sin aportar excesiva información. Se basa en un algoritmo de coste-complejidad (Kim, 1991) y es la opción más recomendable.

En función de lo dicho, vamos a comentar los resultados obtenidos mediante el examen de las tablas y de los árboles que se adjuntan.

### Resultados del rendimiento en *Lengua Española*

Antes de desarrollar el árbol, siempre es importante retener el riesgo asociado al nodo-raíz pues ésta es la varianza total (el riesgo estimado para un árbol con un solo nodo o nivel). En este caso el valor es de 10,5702. Si ahora hacemos que el programa genere el árbol hasta satisfacer un criterio de parada y posteriormente proceda a *podar* los elementos innecesarios, obtenemos (ver figura 3) un modelo con 3 niveles y 13 nodos, 7 de ellos terminales. En este modelo el riesgo estimado es 7,37549.

Estos son los datos para la muestra de aprendizaje, pero para la de validación el riesgo inicial es de 10,8012 y el final de 7,98443. La varianza total es igual a la varianza intra-nodo (error) más la varianza entre-nodo (explicada). La diferencia entre la total (riesgo inicial) y la del error (riesgo final) nos proporciona la varianza explicada. A efectos de evaluación del árbol es más intuitivo a continuación dividir la diferencia entre la

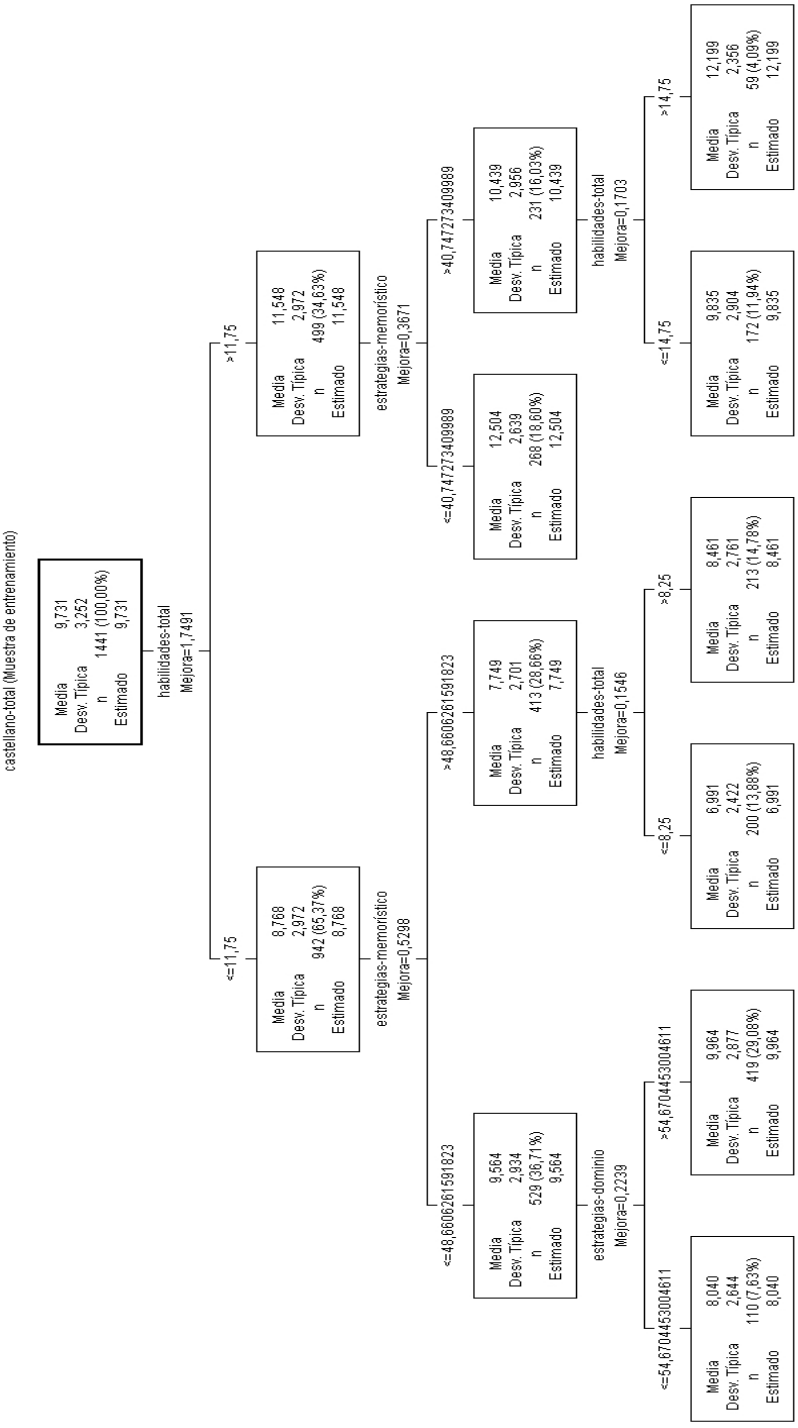


Figura 3
*Árbol* (podado) *de decisión para la variable rendimiento en lengua española*

total, con objeto de expresarla en términos de proporción o porcentuales si el resultado lo multiplicamos por 100.

Con el fin de resumir estos datos y facilitar la evaluación del ajuste de los modelos propuestos hemos elaborado la tabla 3:

TABLA 3

| Lengua española | Riesgo inicial (Varianza total) | Riesgo final (Varianza error) | Diferencia (Varianza explicada) | Proporción de varianza explicada (%) |
|---|---|---|---|---|
| Muestra de aprendizaje | 10,5702 | 7,37549 | 3,19471 | 30,22% |
| **Muestra de validación** | **10,8012** | **7,98443** | **2,81677** | **26,07%** |

Del examen de la misma y del árbol podemos extraer las siguientes conclusiones:

- La muestra de aprendizaje tiende a subestimar los riesgos. Es siempre más adecuado considerar los valores de la muestra de validación para evaluar el modelo.
- De cualquier forma, en este caso, la proporción de variable explicada no llega ni a la tercera parte por lo que en principio habría que concluir que el modelo propuesto no ajusta bien. Esto, sin dejar de ser cierto, no es óbice para que el árbol nos sea de utilidad tanto para el examen de los subgrupos como para la construcción posterior de un modelo paramétrico más detallado.
- Los resultados para el rendimiento en *Lengua Española* nos indican que nos encontramos básicamente con dos variables de segmentación: la puntuación total en *Habilidades metacognitivas* y las *Estrategias memorísticas*. Al final, y segmentando un nodo, aparece también una tercera variable: las *Estrategias-dominio*.
- La segunda variable que segmenta es la de *Estrategias memorísticas*. Y vemos cómo la relación que guarda con el rendimiento es inversa: en todos los pares de nodos, el de la puntuación alta de *Estrategias memorísticas* tiene la media aritmética más baja en rendimiento en *Lengua Española*.
- Estos resultados son coincidentes con los de la regresión por pasos que antes expusimos.
- Pero además la segmentación nos proporciona para cada nivel el *punto de* corte en la variable predictora que establece la partición. Por ejemplo, en el primer nivel los sujetos que obtienen un rendimiento bajo en lengua española (media 8,768) son los que obtienen una puntuación igual o inferior a 11,75 en el total de las habilidades metacognitivas.
- Este primer somero examen apunta en la línea que ya habíamos anticipado: es posible —y útil— emplear las técnicas de segmentación como herramienta exploratoria que permite obtener modelos más finos y parsimoniosos que pueden ser luego empleados con los métodos paramétricos tradicionales.

### Resultados del rendimiento en *Lengua Vasca*

La figura 4 muestra un árbol de 5 niveles con 15 nodos, 8 de ellos terminales. Al igual que en el caso del rendimiento en castellano, los valores de las varianzas se resumen en la tabla 4:

TABLA 4

| Lengua vasca | Riesgo inicial (Varianza total) | Riesgo final (Varianza error) | Diferencia (Varianza explicada) | Proporción de varianza explicada (%) |
|---|---|---|---|---|
| Muestra de aprendizaje | 13,176 | 7,2014 | 5,9746 | 45,34% |
| **Muestra de validación** | **13,4129** | **7,25701** | **6,15589** | **45,89%** |

Si ahora comparamos estos resultados con los de *Lengua Española*, vemos dos diferencias claras. En primer lugar se trata de un árbol más complejo, con más niveles y nodos que en el caso de la *Lengua Española*. Segunda, el porcentaje de la varianza explicada es notablemente mayor en este caso (casi la mitad).

Ambas diferencias se explican con la incorporación al modelo de una nueva variable que antes no aparecía: ***el modelo lingüístico*** en que estudian los sujetos. Es una variable categorial que aparece en primer lugar y luego en varios niveles lo que denota su importancia. De hecho, en su primera aparición (en la primera partición) supone una muy importante mejoría en la disminución del riesgo (3,8386 cuando el total de varianza explicada es de 6,15589, más de la mitad).

Al margen de esto, si observamos el resto del árbol, vemos como las otras variables predictoras siguen siendo las *Habilidades-total* y las *Estrategias memorísticas*. Esta coincidencia nos permite suponer que ambas variables han de ser consideradas como relevantes predictoras del rendimiento en el aprendizaje de lenguas, resultado coincidente con el obtenido en otras evaluaciones (Marchesi y Martín, 2002; Equipo REDES, 2000).

Lo que además ocurre en el caso de la *Lengua Vasca* es que el *modelo* es crucial. Al igual que sucede en el aprendizaje de otras lenguas, existe una clara diferencia entre los sujetos que estudian la lengua vasca sólo como asignatura, frente aquellos modelos en que —total o parcialmente— se estudia ***en*** dicha lengua. En lo relativo al español, el modelo lingüístico no resulta ser una variable relevante porque dada la situación socio-lingüística de la Comunidad Autónoma Vasca, todos los sujetos aprenden el español al margen de que sea considerado como asignatura o como lengua vehicular.

Estos resultados son coincidentes con las investigaciones realizadas para comprobar la eficacia de los mencionados modelos (Etxeberria Balerdi, 1999; Idiazabal & Kaifer, 1994; Lukas, 1994). A raíz de estas investigaciones se ha creado un *corpus* teórico acerca

de los resultados que obtienen los estudiantes en lengua española y en lengua vasca tras haber sido escolarizados en los distintos modelos lingüísticos. Las conclusiones más relevantes que se han obtenido son las siguientes:

- El rendimiento en *Lengua Vasca* está influído por diversos factores, sin embargo, el factor más relevante es el modelo lingüístico. De tal forma que se puede afirmar que los estudiantes del modelo D son los que obtienen puntuaciones significativamente más altas que el resto de los estudiantes. A continuación se encuentran los estudiantes escolarizados en el modelo B y por último y a una mayor distancia los del modelo lingüístico A.
- En nuestro caso, en el árbol de la figura 4 podemos ver cómo en primer lugar se segmenta distinguiendo entre el modelo A por una parte y el B y D por otra. En esta primera segmentación los sujetos escolarizados en el modelo A obtienen una nota media en lengua vasca de 5,078 mientras que los de los modelos B y D alcanzan un 9,712.
- El rendimiento en *Lengua Española* por el contrario, no está condicionado por el modelo lingüístico seguido durante la escolarización. Independientemente del modelo lingüístico en el que ha sido escolarizado el estudiante, su rendimiento en *Lengua Española* no varía. Son otros los factores que determinan un mayor o menor rendimiento en dicha materia.

Como vemos, estos resultados son básicamente coincidentes con los que hemos obtenido empleando tanto las técnicas de análisis factorial, las de regresión, y, por último, las de segmentación.

En nuestra opinión, la ventaja de estas últimas es que nos permiten examinar la estructura introduciendo en el modelo todas las variables que se estime oportuno independientemente de su nivel de medida y sin tener que someterlas a priori a ningún tipo de recodificación.

## CONCLUSIONES

Con respecto al **primer objetivo** planteado, es decir, la propuesta de un modelo predictivo del rendimiento en las materias de lengua española y lengua vasca, las principales conclusiones son las siguientes:

- Los resultados de la triangulación efectuada permiten verificar que los de la segmentación son coincidentes con los de las técnicas clásicas de regresión y de reducción de la dimensionalidad.
- Las variables que se han comportado como mejores predictoras del rendimiento en español son, por este orden, la puntuación total en las *habilidades metacognitivas* y el empleo de *estrategias de aprendizaje basadas en la memorización*, ésta última guardando una relación inversa. En un segundo nivel, aparece también como predictora el empleo de *estrategias de dominio.* Estos resultados coinciden con los obtenidos con otras muestras (Marchesi y Martín, 2002).

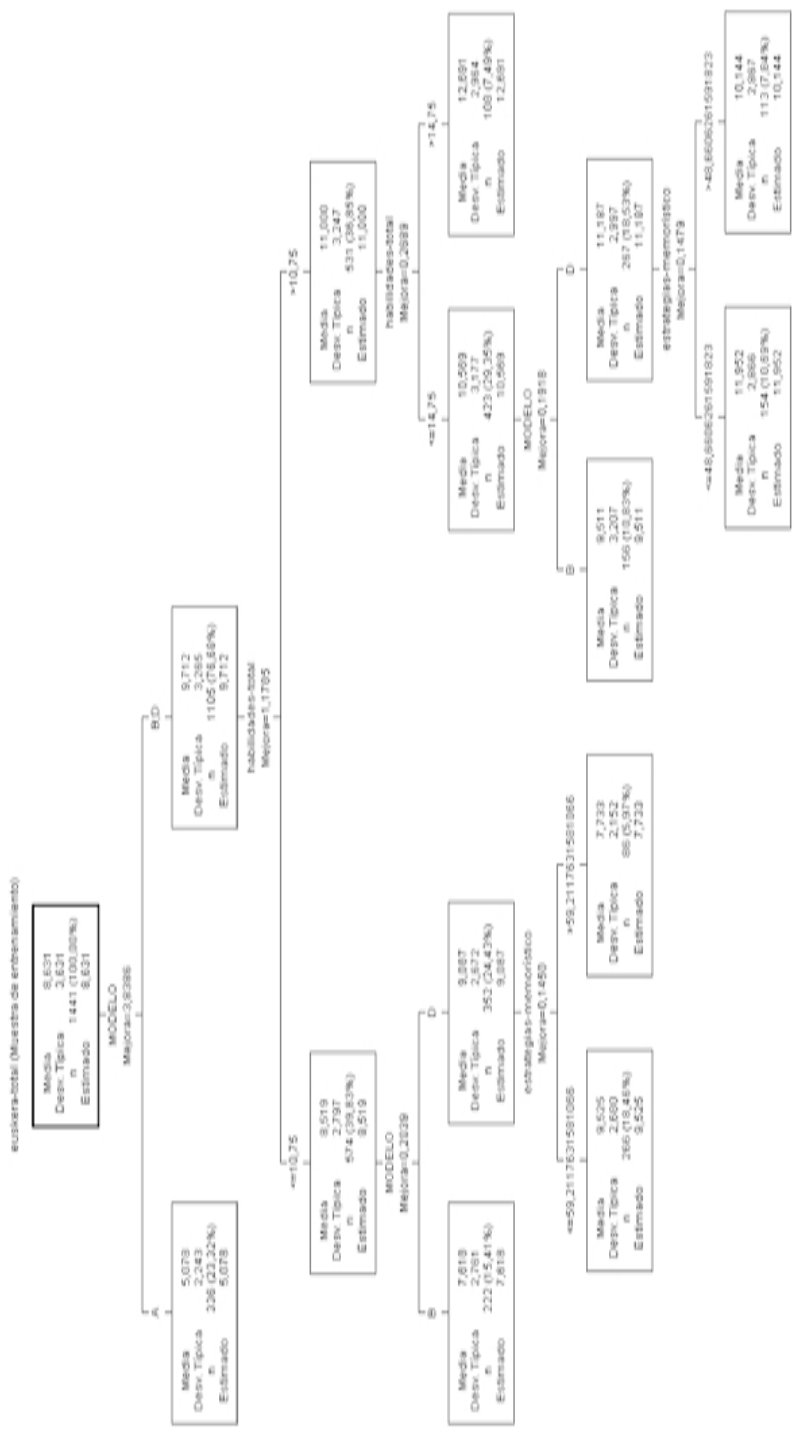


Figura 4
*Árbol* (podado) *de decisión para la variable rendimiento en lengua vasca*

- En el caso del rendimiento en lengua vasca, a estas variables se añade en lugar preferente el *modelo lingüístico* en el que los estudiantes son escolarizados. Aquí los resultados también confirman la tendencia encontrada en investigaciones previas (Etxeberria Balerdi, 1999; Idiazabal & Kaifer, 1994; Lukas, 1994).

En lo tocante al **segundo objetivo** formulado consistente en examinar las posibilidades que las técnicas de segmentación ofrecen en este campo podemos concluir afirmando que:

- Las técnicas de segmentación constituyen una herramienta exploratoria que puede resultar de gran utilidad en este tipo de problemas habida cuenta de que los árboles de decisión ofrecen la ventaja de poder operar simultáneamente con todo tipo de variables.
- Además facilitan la identificación de la interacción, pues las variables predictoras se utilizan unas en relación con otras permitiendo también la caracterización de subpoblaciones.
- Las tablas y gráficos que se emplean para mostrar los resultados son de sencilla interpretación lo que es muy importante de cara a la presentación y comunicación de resultados a audiencias no expertas.
- En definitiva, se trata de una útil herramienta exploratoria que permite obtener pautas para diseñar modelos más depurados de cara a su análisis posterior mediante técnicas paramétricas. Éste era el objetivo metodológico en la fase previa, la exploración. No es que se haya tratado de eludir el aspecto inferencial en cuanto a la investigación de un modelo. Una vez allanado el terreno, tal quehacer se abordará por medio de los modelos jerárquicos lineales. Pero se trata de un objetivo que aquí no tiene cabida.

## REFERENCIAS BIBLIOGRÁFICAS

Biggs, D.B. de Ville & Suen, E. (1991). «A method of choosing multiway partitions for classification and decision trees». *Journal of Applied Statistics*. N18. 49-62.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). *Classification and regression trees*. Belmont (California): Wadsworth.

Equipo REDES (1999). «Una red de evaluación de centros de educación secundaria». *Infancia y Aprendizaje*. N 85. 59-73.

Equipo REDES (2000). «Variables predictoras de la calidad de la educación secundaria». *Revista de Educación*, núm. 323, pp. 395-420.

Etxeberria Balerdi, F. (1999). *Bilingüismo y educación en el país del euskara*. San Sebastián: Erein.

Everett, P. and others. (1997). «Presentation of Social and Academic Factors that encourage persistence in Secondary Schools in rural, low socioeconomic areas of two selected southeastern states». Paper presented at the *Annual Meeting of the American Educational Research Association*. AERA. Chicago, Illinois.

Forthofer, M.; Bryant, C. (2000): «Using audience-segmentation techniques to tailor health behavior change strategies». *American Journal of Health Behavior*. Vol 24 (1), pp. 36-43.

Godley, S.; Fiedler, e.; Funf, R. (1998). «Consumer satisfaction of parents and their children with child/adolescent mental health services». *Evaluation and program planning*. No 21, 1, pp. 31-45.

Hambleton, R.K. (1996). Adaptación de tests para su uso en diferentes idiomas y culturas: fuentes de error, posibles soluciones y directrices prácticas. En J. Muñiz (Coor.). *Psicometría*. Madrid: Universitas. 207-238.

Henry, G.T. (1993). «Using graphical displays for evaluation data». *Evaluation Review*. 17, 60-78.

Henry, G.T. (1998). Graphing Data, en BICKMAN, L. y ROG, D. J. (Eds.) *Handbook of Applied Social Research Methods*. Thousand Oaks, CA: Sage, 527-556.

Idiazabal, I. & Kaifer, A. (Ed.). (1994). *Eficacia educativa y enseñanza bilingüe en el País Vasco*. Vitoria: I.V.A.P.

Jornet Meliá, J.M. & Suárez Rodríguez, J.M. (1996). «Pruebas estandarizadas y evaluación del rendimiento: usos y características métricas». *Revista de Investigación Educativa. V14. N2.* 141-163.

Kass, G. (1980). «An exploratory technique for investigating large quantities of categorical data». *Applied Statistics. V29. N2.* 119-127.

Kim, S.H. (1991). *An extension of CART's Pruning Algorithm.* Program Statistics Research Technical Report No 91-11. Educational Testing Service, Princeton, New Jersey.

Lizasoain, L. y Joaristi, L. (2000). «El análisis de datos en la evaluación de programas educativos». *Revista de Investigación Educativa*. Vol. 18, Nº 2, pp. 357-379.

Loh, W.Y.; Vanichsetakul, N. (1988). «Tree-structured classification via generalized

Lukas Mujika, J.F. (1994). *Trebetasun eta errendimendu matematikoa testuinguru elebidunean.* Lejona: Servicio Editorial de la Universidad del País Vasco.

Marchesi, A. y Martín, E. (comp.) (2002). *Evaluación de la educación secundaria. Fotografía de una etapa polémica*. Madrid. SM.

Patton, M.Q. (1997). *Utilization-Focused Evaluation. The New Century Text (3rd ed.).* Thousand Oaks, CA: Sage.

Repetto, E. y otros (1994). «Últimas aportaciones en la evaluación del programa de orientación metacognitiva de la comprensión lectora». *Revista de Investigación Educativa*, 23, 314-323.

Santiago, C. & Lukas, J.F. (2000). «Evaluación externa de centros en la Comunidad Autónoma Vasca». Ponencia presentada en las *I Jornadas sobre Medición y Evaluación Educativas: Estándares e Indicadores para analizar la realidad educativa.* Valencia, 8,9 y 10 de marzo de 2000. En prensa.

SPSS Inc. (1999). *Answer Tree*. SPSS Inc., Chicago.

Fecha de recepción: 30 de mayo de 2001.
Fecha de aceptación: 12 de septiembre de 2002.