

Predicción del rendimiento en una asignatura empleando la regresión logística ordinal

Predicting Performance in a Subject Using Ordinal Logistic Regression

Prevendo desempenho em um assunto por meio de regressão logística ordinal

Jobany J. Heredia R.,^a Aida G. Rodríguez H., José A. Vilalta A.

^aDepartamento de Ingeniería Industrial, Facultad de Ingeniería Industrial. Instituto Superior Politécnico “José Antonio Echeverría”, La Habana, Cuba. Correo electrónico: jherediar@ind.cujae.edu.cu

RESUMEN

En las materias donde el índice de fracaso es considerable, es fundamental que el profesor posea información relevante sobre sus alumnos para desarrollar un tratamiento diferenciado en la enseñanza. En el trabajo se emplea la regresión logística ordinal para construir una ecuación que relacione la calificación en la asignatura Modelos Probabilísticos de los Procesos (MPP), la cual se imparte en segundo año de la carrera de Ingeniería Industrial, con sus resultados en primer año. Con los datos de 274 estudiantes pertenecientes a dos cursos académicos distintos, se obtuvo como mejor ecuación la que relaciona la evaluación en MPP con el promedio del alumno en las asignaturas de ciencia que recibe en primer año. Las estimaciones de probabilidad conseguidas con este modelo se usaron como base para el desarrollo de un método que permitió predecir el estado en la asignatura de los alumnos que la recibieron a inicios del curso 2011-2012.

Palabras clave: rendimiento estudiantil, predicción, regresión logística ordinal.

ABSTRACT

Subjects with important failure rates demand teachers having relevant information about their students to accomplish a differentiated teaching process. Ordinal logistic regression is used to build an equation that relates students' marks in Probabilistic Models of Processes (MPP according to its acronym in Spanish), which is a subject taught in the second year of Industrial Engineering studies, with first year results. Data from 274 students belonging to two different academic years were used to achieve as the best equation, the one that relates MPP grade with the average mark of science subjects studied in the first year. Probability estimations obtained with this model were used as a starting point to develop a method that allowed the prediction of MPP condition for those students who took this subject in the academic year 2011-2012.

Key words: academic performance, prediction, ordinal logistic regression.

RESUMO

Em conteúdos onde o índice de fracasso é significativo, é essencial que o professor tenha informações relevantes sobre seus alunos para desenvolver uma prática de ensino diferenciada. No trabalho, emprega-se a regressão logística ordinal para construir uma equação que relaciona a qualificação à disciplina, Modelos Probabilísticos de Processos (MPP), ministrada no segundo ano de Engenharia Industrial, a partir dos resultados do primeiro ano. Partindo dos dados de 274 estudantes provenientes de dois anos letivos diferentes, concretizou-se como melhor equação aquela que relaciona a avaliação em MPP com a média do estudante das disciplinas de ciências recebidas no primeiro ano. As estimativas de probabilidade obtidas a partir de esse modelo foram utilizadas como base para o desenvolvimento de um método que nos permitiu prever o estado na disciplina cursada por estudantes no início dos anos letivos 2011-2012.

Palavras chave: desempenho estudiantil, previsão, regressão logística ordinal.

1. INTRODUCCIÓN

La capacidad de las organizaciones educativas para desarrollar un tratamiento diferenciado de sus educandos es reconocida como una característica fundamental para alcanzar la excelencia en la educación superior (Organización Internacional de Normalización, 2002). El logro de acciones educativas diferenciadas requiere que los docentes conozcan las características de sus alumnos para que puedan guiarlos, adecuándose a sus capacidades y motivaciones, y esto, a su vez, exige perfeccionar la gestión en las instituciones de enseñanza superior, de manera que sea posible brindar a estos docentes informaciones más completas y oportunas sobre sus estudiantes (Horrutiner, 2007).

La aplicación de técnicas estadísticas favorece la comprensión de la variabilidad relacionada con las razones de deserción, la satisfacción y los indicadores de desempeño del educando, no solamente porque permite medir y cuantificar estas variables, sino también por la posibilidad que brinda para obtener predicciones de las mismas (Organización Internacional de Normalización, 2002). Esta situación, unida a la ventaja que representa para estos fines el gran desarrollo que han experimentado los sistemas de información universitarios en cuanto al volumen de datos viables de ser analizados, promueve el despliegue del análisis de datos para el descubrimiento y la gestión del conocimiento asociado al proceso de enseñanza (Luan, 2002). De hecho, a partir del año 2000 en la gestión de las instituciones de educación superior ha ganado auge el empleo de técnicas estadísticas y de minería de datos (Delavari & Beikzadeh, 2008).

Dentro de estas aplicaciones, las más frecuentes son las que utilizan técnicas matemáticas para predecir el rendimiento estudiantil, ya sea en una materia en particular o en un período lectivo (Luan, 2002). Las técnicas más empleadas con este fin han sido los árboles de clasificación, las redes neuronales, la regresión logística y el análisis discriminante (Luan, 2002). Generalmente, estas técnicas se han aplicado para predecir simplemente los estados de “éxito” o “fracaso”, pero en las universidades cubanas las calificaciones de las materias se emiten en una escala ordinal, variando desde la más baja categoría de “mal”, la cual equivaldría a “fracaso”, hasta la más alta de “muy bien”. Por lo tanto, para considerar en este contexto un estudio de desempeño docente con toda la información que se deriva de los resultados académicos, debe emplearse un método que tenga en cuenta la naturaleza ordinal de estos resultados, siendo la regresión logística ordinal una técnica que se ajusta a esta exigencia.

La regresión logística ha tenido un extendido uso por su capacidad para tratar variables independientes, tanto numéricas como categóricas, y por la utilidad de la información que se deriva del análisis del denominado “*odds ratio*” (Agresti, 1990; Hosmer & Lemeshow, 2000). Los modelos logísticos son adecuados para situaciones donde se quiere explicar la probabilidad “*p*” de ocurrencia de un evento de interés por medio de los valores de ciertas variables independientes (Llinás, 2006). Cuando los eventos cuya probabilidad se desea explicar corresponden a variables categóricas ordinales, es decir, aquellas cuyos valores no sólo diferencian a los individuos, sino que también permiten establecer un orden entre estos, es recomendable el empleo de la regresión logística ordinal, pues con la aplicación de una técnica de predicción que solamente considere la variable dependiente como nominal (por ejemplo, la regresión logística multinomial o el análisis discriminante), no se captaría cabalmente la influencia de las variables explicativas sobre la variable dependiente al no considerar la información acerca de las diferencias de orden entre las categorías de esta última (McCullagh, 1980; Agresti, 1990; Hosmer & Lemeshow, 2000).

En este trabajo se aplica la regresión logística ordinal con el objetivo de encontrar un modelo explicativo para las calificaciones de los alumnos en la asignatura Modelos Probabilísticos de los Procesos (materia que se imparte en segundo año de la carrera de Ingeniería Industrial del Instituto Superior Politécnico “José Antonio Echeverría” (ISPJAE)), empleando los datos docentes sobre los mismos que se registra en primer año. Este modelo desea usarse para obtener, antes de comenzar la impartición de la materia, una predicción del resultado que alcanzará cada alumno. Esta información facilitaría el desarrollo de acciones educativas diferenciadas y con carácter proactivo con los estudiantes.

El primer factor que motiva este estudio es el hecho de que los autores del trabajo son profesores de la asignatura Modelos Probabilísticos de los Procesos y, por ende, se benefician directamente con sus resultados. Además, esta materia, junto a Física 2 y Matemática 3, son las de más bajo índice de promoción en el semestre donde se imparten. Con esta investigación también se pretende dar un uso más amplio a los datos docentes sobre los alumnos que se acumulan en el sistema de información con que cuenta la facultad, pues actualmente sólo se emplean para clasificar el desempeño del alumno al final de cada semestre.

El artículo comienza haciendo un breve esbozo teórico sobre la regresión logística ordinal en donde se comentan las estadísticas, las distribuciones de probabilidad y los requisitos asociados a esta técnica. Después se aplica esta variante de regresión para la obtención del modelo deseado, y, posteriormente, se analiza la utilidad de los resultados obtenidos. Éste es un estudio con carácter exploratorio, pues, aunque en el ISPJAE se han desarrollado estudios similares (Alfonso, 2008; Brito, 2008), no existen precedentes de este tipo de investigación en la carrera de Ingeniería Industrial. Los resultados se obtienen empleando como software el Minitab en su versión 15.

2. REGRESIÓN LOGÍSTICA ORDINAL (RLO). CONSIDERACIONES TEÓRICAS.

La regresión logística en su forma más simple, es decir, con una respuesta binaria, propone que el logaritmo de la “razón de probabilidad” (*odds* según su denominación en inglés), entendida como el cociente entre la probabilidad de éxito y la de fracaso en un ensayo de Bernoulli, es igual a una función lineal en los parámetros, denominada usualmente predictora lineal (Ponsot *et al.*, 2009: 158). En el caso particular de la RLO se utiliza una función de unión para relacionar de forma lineal a las variables explicativas con la razón de probabilidad entre la probabilidad acumulada hasta la categoría i de la variable ordinal, y la probabilidad que la variable tome un valor mayor que la categoría i (Agresti, 1990; Hosmer & Lemeshow, 2000).

En la RLO las funciones de unión más empleadas son la Logit y la Cloglog (Agresti, 1990). McCullagh (1980) plantea que la unión con Logit es más adecuada para analizar datos ordinales, cuya distribución de frecuencia es uniforme a lo largo de todas las categorías, mientras que la unión Cloglog es preferible para analizar datos categóricos, cuyas categorías de mayor valor son las más probables. En el caso de estudio que se muestra en este trabajo, los valores de la variable ordinal representan las calificaciones en una asignatura cuya complejidad es valorada por los educandos como de media a alta, por ende no ocurre que los mayores valores (mejores calificaciones) sean los más probables, siendo plausible considerar la unión Logit como las más satisfactoria para este caso.

La expresión de la función Logit para la RLO es la siguiente:

$$\ln(O_i) = \alpha_i + \beta X \quad (1)$$

En esta ecuación, O_i es la “razón de probabilidad” (odds) asociada a la categoría i de la variable dependiente, siendo la expresión de esta razón:

$$O_i = \frac{P(\text{valor sea } \leq \text{categoría } i / \text{valores de } X)}{P(\text{valor sea } > \text{categoría } i / \text{valores de } X)} \quad (2)$$

Lo que es lo mismo que:

$$O_i = \frac{P(\text{valor sea } \leq \text{categoría } i / \text{valores de } X)}{1 - P(\text{valor sea } \leq \text{categoría } i / \text{valores de } X)} \quad (3)$$

Con el término “valor” en las expresiones (2) y (3) se hace referencia a cualquier valor de la variable dependiente. Como se observa, las probabilidades de las expresiones (2) y (3) son condicionales, es decir, dados los valores de la (s) variable(s) independiente(s). En la ecuación (1), α_i es el intercepto asociado a la ecuación que modela la razón de probabilidad de la categoría i , y β es el coeficiente de la ecuación de regresión. Si existen p variables independientes, existen p coeficientes, y βX se reemplaza por la combinación lineal entre $\beta X_1 + \beta X_2 + \dots + \beta X_p$. Estos coeficientes cuantifican el efecto de las variables independientes sobre el logaritmo de la razón de probabilidad.

Si la variable dependiente tiene k categorías, existen $k-1$ ecuaciones ya que a la categoría mayor no se asocia odds, al ser la probabilidad acumulada hasta ésta igual a uno. El modelo de regresión ordinal descrito anteriormente es denominado “modelo logit acumulado”, ya que es construido basándose en las probabilidades acumuladas de la variable respuesta, dados los valores de las variables explicativas (Hosmer & Lemeshow, 2000). También es denominado “modelo de razón de probabilidad proporcional”, ya que los coeficientes de regresión son independientes de las categorías de la variable dependiente, siendo los mismos en las $k-1$ ecuaciones que se forman para las categorías. Esto implica asumir que la relación entre las variables explicativas y la variable dependiente ordinal es independiente de las categorías de esta última, y, por tanto, que los cambios en las variables explicativas provocan el mismo cambio en la razón de probabilidad acumulada de todas las categorías (McCullagh, 1980; Agresti, 1990; Hosmer & Lemeshow, 2000). Este supuesto debe comprobarse a través del test de líneas paralelas o de forma gráfica (Hosmer & Lemeshow, 2000), por tanto existen ecuaciones $k-1$ con los mismos coeficientes acompañando a las variables explicativas y que sólo se diferencian en el valor del intercepto.

Para estimar los coeficientes de la ecuación de regresión se utilizan diversos procedimientos, siendo la estimación de máxima verosimilitud el más empleado (Agresti, 1990). Una vez obtenida la ecuación, un primer aspecto importante es comprobar la significancia global de la ecuación, lo que significa evaluar si en conjunto las variables regresoras tienen un efecto estadísticamente significativo sobre la variable dependiente. Con este fin se emplea el estadístico G , calculado según la fórmula:

$$G = -2 \ln \left(\frac{\text{verosimilitud de la muestra sin variables explicativas}}{\text{verosimilitud con la ecuación que incluye las variables}} \right) \quad (4)$$

Este estadístico sigue distribución chi-cuadrado, con tantos grados de libertad como variables independientes existan (McCullagh, 1980). Cuando las predicciones de la variable dependiente que se hacen con el modelo que incluye todas las variables independientes superan las predicciones que se realizan sin considerar éstas, el valor de G tiende a ser grande y conlleva a concluir que al menos una de las variables regresoras tiene efecto sobre la variable dependiente, y, por lo tanto, que la probabilidad de ocurrencia de los valores que representan esta variable varía para alguna de las combinaciones de valores de las variables independientes (Agresti, 1990; Hosmer & Lemeshow, 2000). La significancia individual de cada variable independiente, generalmente, se analiza a través de la prueba de Wald, la cual se basa en la significancia del coeficiente estimado para cada variable (Agresti, 1990). El estadístico utilizado es el siguiente:

$$ZWald = b_j ES(b_j) \quad (5)$$

Donde b_j es el coeficiente de regresión estimado para la variable independiente j .

Bajo la hipótesis de que el coeficiente poblacional $\beta_j = 0$ para la variable j , la razón entre la estimación de este coeficiente (b_j) y el error estándar de esta estimación [$ES(b_j)$], sigue una distribución normal estándar. Mayores valores de este estadígrafo indican que el coeficiente β_j es distinto de cero, y, por ende, que la variable independiente tiene efecto sobre la probabilidad de ocurrencia de los valores de la variable dependiente.

También es necesario conocer si el modelo que se obtiene presenta buen ajuste. Usualmente, para evaluar el ajuste se construye una tabla de contingencia, cuyas filas representan los valores de la variable objeto de estudio y cuyas columnas representan las posibles combinaciones de valores de las variables independientes (Hosmer *et al.*, 1997). El empleo de una prueba de bondad de ajuste permite saber si la frecuencia predicha para las combinaciones según el modelo difiere significativamente de la frecuencia con la cual ocurren realmente los valores en estas combinaciones, lo cual evidencia falta de ajuste (Hosmer *et al.*, 1997; Hosmer & Lemeshow, 2000). Para comparar estas frecuencias generalmente se calcula el estadístico chi cuadrado de Pearson según la fórmula (Hosmer *et al.*, 1997):

$$\chi^2 = \sum_{i=1}^k \sum_{l=1}^m \frac{(y_{il} - m_{il})^2}{m_{il}} \quad (6)$$

En la expresión (6) se considera que la variable dependiente tiene k categorías y que se forman m combinaciones de valores con las variables explicativas, de manera que:

- y_{il} es la frecuencia observada de la i -ésima categoría de la variable dependiente en la l -ésima combinación de valores de las variables explicativas.
- p_{il} es la probabilidad estimada con el modelo para la i -ésima categoría de la variable dependiente en la l -ésima combinación de valores de las variables independientes.

• m_l es la cantidad de elementos en la l -ésima combinación de valores de las variables explicativas.

Mientras mayor es el valor del estadístico χ^2 mayor sospecha de falta de ajuste.

Si finalmente se concluye la existencia de relación entre las variables explicativas y la dependiente, y si la ecuación lograda presenta buen ajuste, entonces se pueden hacer otros análisis, por ejemplo, para obtener la razón de probabilidad acumulada de la categoría i de la variable dependiente para determinados valores de las independientes, se despeja esta razón de la función logarítmica de forma que:

$$P(\text{valor sea } \leq \text{categoría } i / \text{valores de } X) - P(\text{valor sea } > \text{categoría } i / \text{valores de } X) = e^{a_i + \beta X} \quad (7)$$

Con el término “valor” en la expresión (7) se hace referencia a cualquier valor de la variable dependiente. De (7) se deriva que:

$$P(\text{valor sea } \leq \text{categoría } i / \text{valores de } x) = (e^{a_i + \beta X}) / (1 + e^{a_i + \beta X}) \quad (8)$$

Y de (8) se deduce que:

$$P(\text{valor sea } = \text{categoría } i / \text{valores de } x) = P(\text{valor sea } \leq \text{categoría } i) - P(\text{valor sea } \leq \text{categoría } i - 1) \quad (9)$$

La expresión (9) es de suma utilidad, pues posibilita estimar a través de la ecuación obtenida y dado un conjunto de valores de las variables regresoras, la probabilidad que la dependiente tome cada uno de sus valores.

También suele calcularse el ratio de la razón de probabilidad (*odds ratio*) que provoca el cambio en cada una de las variables independientes. El *odds ratio* de la variable independiente x evalúa la relación entre la razón de probabilidad asociada a la categoría i cuando $x = x_2$, y la razón de probabilidad asociada a la categoría i cuando $x = x_1$. Numéricamente sería:

$$\text{odds ratio} = \frac{P(Y \leq i | X = X_2) / P(Y \geq i | X = X_2)}{P(Y \leq i | X = X_1) / P(Y \geq i | X = X_1)} \quad (10)$$

Como el efecto que tiene una determinada variable predictora es el mismo para todas las categorías de la variable dependiente, para cada variable independiente se determina un solo *odds ratio*. El *odds ratio* es utilizado para interpretar el efecto de las variables explicativas sobre la variable objeto de estudio. Si éste es igual a uno, indica que la variable predictora no tiene efecto. Si es menor que uno, lo cual sucede cuando el coeficiente de la variable regresora es negativo, indica que, si las otras variables explicativas permanecen constantes, los cambios en la variable explicativa analizada incrementan la probabilidad de obtener categorías de mayor valor en la variable objeto de estudio (Agresti, 1990; Hosmer & Lemeshow, 2000). Valores de *odds ratio* mayores que uno muestran que las variaciones en la variable independiente disminuyen la probabilidad de obtener categorías de mayor valor de la dependiente.

3. MÉTODO

Se aplicó la regresión logística ordinal para obtener un modelo que permitiera pronosticar los resultados en la asignatura Modelos Probabilísticos de los Procesos (MPP), la cual se imparte en el primer semestre de segundo año de la carrera de Ingeniería Industrial. La variable dependiente es la calificación en esta asignatura, variable cuya escala es ordinal, pues sus posibles valores son: “mala”, regular”, “buena” y “muy buena”. En las universidades cubanas, para facilitar el procesamiento de los datos, se codifican estos valores registrando como “2” las malas calificaciones, “3” las regulares, “4” las buenas y “5” las muy buenas. La calificación de “2” implica que el estudiante está reprobado.

Como variables independientes se utilizaron diferentes variables relacionadas con los resultados que obtiene el alumno en su primer año de estudio. En diversas investigaciones se ha encontrado la existencia de relación entre las calificaciones que obtienen los alumnos en las materias de segundo año con las evaluaciones en las materias de primero (Zúñica *et al.*, 2005; Delavari & Beikzadeh, 2008). Como se ha comentado, este es un estudio exploratorio, por lo que, a pesar de que es conocida la existencia de otros factores, además de las calificaciones en materias precedentes, que influyen en el desempeño estudiantil, no se consideran en esta investigación. No obstante, esto no le resta valor al análisis pues, como se acotó en la Introducción, se está proponiendo un nuevo uso a los datos docentes de los alumnos de primer año, los cuales actualmente tienen un uso limitado. Además, de encontrar ventajas al aplicar la RLO de la forma que se propone, puede extenderse esta aplicación, incluyendo otros factores explicativos.

Inicialmente se proyectó considerar como única variable explicativa el rendimiento global del estudiante en primer año, pero, considerando que este rendimiento está formado por el desempeño en una serie de materias, se decidió hacer un análisis más detallado para evitar asignar al rendimiento global, un efecto que pudiera ser solamente de un subgrupo de asignaturas. Por esta razón se decidió formar subgrupos con las materias de primer año, de manera que en un subgrupo se agruparan asignaturas semejantes en cuanto a sus contenidos, definiendo como variable independiente el promedio del estudiante en cada uno de los conjuntos de materias semejantes. De esta forma se establecieron como variables independientes las siguientes:

- Promedio del estudiante en las asignaturas de ciencia, las cuales incluyen, Matemática 1 y 2, Álgebra, Física 1 y Química.
- Promedio del estudiante en las asignaturas de dibujo, las cuales incluyen Dibujo Básico y Dibujo Aplicado.
- Promedio del estudiante en las asignaturas de inglés, las cuales incluyen Inglés 1 y 2.
- Promedio del estudiante en las asignaturas más técnicas, las cuales incluyen Introducción a la Ingeniería e Introducción a la Ingeniería Industrial.
- Promedio del estudiante en las asignaturas de ciencias sociales, las cuales incluyen Historia, Filosofía y Economía Política.
- La calificación en Introducción a la Informática, la cual se decidió no agrupar con ninguna de las materias del año.

Todas las variables explicativas son consideradas numéricas, a excepción de la calificación en Introducción a la Informática, la cual se definió como una variable categórica (ordinal en este caso), cuyas categorías son los cuatro posibles valores que puede tomar esta calificación, es decir, “mal”, “regular”, “bien” y “muy bien”.

Para construir el modelo se emplearon los datos de los 130 estudiantes de primer año en el curso 2008-2009 y de estos mismos estudiantes en segundo año (curso 2009-2010);¹ además, los 144 datos de los estudiantes de primer año en el curso 2009-2010 y de estos mismos alumnos en segundo año (curso 2010-2011).² Estos dos conjuntos de datos constituyen la muestra que se utiliza para ajustar el modelo. Después se aplicó la regresión logística, de forma individual, con cada uno de los dos conjuntos de datos que han sido comentados y, posteriormente, con ambos conjuntos agrupados. Luego de analizar las semejanzas y diferencias en los resultados de estos tres casos, se obtuvo un modelo (empleando ambos conjuntos unidos), cuya efectividad se evaluó de la siguiente manera:

- Se evaluaron los valores de las variables independientes correspondientes, usando los resultados de los 145 alumnos de primer año en el curso 2010-2011.
- Con esta evaluación se obtuvo una predicción del desempeño de estos estudiantes en MPP para el curso 2011-2012.
- Posteriormente, se compararon las predicciones realizadas con los resultados reales obtenidos por estos alumnos y así se valoró la calidad del modelo.

Por lo tanto, los datos de los alumnos que en el curso 2010-2011 cursaban el primer año y en el curso 2011-2012 cursaban su segundo año, constituyeron la muestra que se empleó para evaluar la efectividad del modelo, siendo diferente a las muestras que se emplean para ajustarlo, que es lo recomendado.

4. RESULTADOS

Antes de analizar los resultados conseguidos al aplicar la RLO con el primer conjunto de datos, en la tabla 1 se muestra el cálculo de la media y la desviación estándar de cada una de las variables independientes cuantitativas, diferenciando estos estadísticos para cada posible resultado de la variable dependiente. Cuando se analiza la tabla 1, se evidencia que la calificación promedio de todos los subconjuntos de materias tiende a ir aumentando a medida que va aumentando la calificación en MPP, tendencia que es más explícita para el caso del promedio en las ciencias y el promedio en Dibujo.

Tabla 1. Estadísticas descriptivas para variables cuantitativas. Primer conjunto

Calificación MPP		Promedio ciencias	Promedio dibujo	Promedio inglés	Promedio técnicas	Promedio ciencias sociales
2	Media	3,3	4,07	4,05	4,09	4,47
	Desviación Estándar	0,61	0,7	0,45	0,62	0,36
3	Media	3,74	4,3	4,1	4,27	4,58
	Desviación Estándar	0,41	0,61	0,3	0,75	0,22

¹ Al conjunto de datos asociados a estos estudiantes se les denominará en el trabajo "primer conjunto de datos".

² Al conjunto de datos asociados a estos estudiantes se les denominará en el trabajo "segundo conjunto de datos".

4	Media	3,92	4,53	4,3	4,6	4,72
	Desviación Estándar	0,53	0,59	0,43	0,5	0,27
5	Media	4,44	4,88	4,55	4,72	4,93
	Desviación estándar	0,41	0,53	0,52	0,36	0,15

Al aplicar la RLO con el primer conjunto de datos los resultados fueron los siguientes:³

Tabla 2. RLO aplicada al primer conjunto de datos

Link Function: Logit								
Logistic Regression Table								
Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI		
						Lower	Upper	
Const(1)	20,7806	4,54824	4,57	0,000				
Const(2)	22,0610	4,61055	4,78	0,000				
Const(3)	23,5689	4,69558	5,02	0,000				
Prom Cien	-2,07584	0,483476	-4,29	0,000	0,13	0,05	0,32	
Prom CienSoc	-1,79112	0,868598	-2,06	0,039	0,17	0,03	0,92	
Prom Dib	-0,850120	0,409619	-2,08	0,038	0,43	0,19	0,95	
Prom Tecni	-0,415303	0,400713	-1,04	0,300	0,66	0,30	1,45	
Prom Ingles	-0,132635	0,530109	-0,25	0,802	0,88	0,31	2,48	
ININF								
3	1,22156	1,84681	0,66	0,508	3,39	0,09	126,63	
4	1,25859	1,82603	0,69	0,491	3,52	0,10	126,17	
5	2,04286	1,83577	1,11	0,266	7,71	0,21	281,74	
Log-Likelihood = -94,740								
Test that all slopes are zero: G = 61,999, DF = 8, P-Value = 0,000								
Goodness-of-Fit Tests								
Method	Chi-Square	DF	P					
Pearson	291,123	364	0,998					
Deviance	181,163	364	1,000					

Analizando el valor tan pequeño del valor p de la prueba de la significancia global (Test that all slopes are zero), se rechaza la hipótesis de que los coeficientes de todas las variables explicativas son cero, concluyendo que al menos una de las variable regresoras tiene efecto sobre la calificación en MPP, y, por lo tanto, que la probabilidad de ocurrencia de los valores que representan esta calificación varía para alguna de las combinaciones de valores de las variables independientes. Las pruebas de bondad de ajuste (Goodness of fit tests) no dan evidencia para considerar que hay falta de ajuste (se acepta la hipótesis nula pues el valor p es mayor que el nivel de significancia), lo que significa que las probabilidades de ocurrencia de los valores de la variable dependiente que se estiman

³ En todos los análisis de este trabajo se emplea un 5% como nivel de significancia.

según el modelo para las diferentes combinaciones de las independientes, no divergen significativamente de la frecuencia con la cual ocurren en la muestra los valores de la variable dependiente para estas combinaciones.

En la cuarta y quinta columnas de la tabla de regresión logística se muestra el valor Z del estadístico de Wald asociado a cada variable independiente y su correspondiente valor p.⁴ A través de estos se puede analizar la significancia individual de las variables explicativas. El análisis de este primer conjunto de datos evidencia que sólo el promedio del estudiante en las ciencias, el promedio en las ciencias sociales y el promedio en Dibujo, tuvieron una relación significativa con la calificación en MPP, pues solamente los valores p asociados a la significancia de los coeficientes de estas variables son menores que 0,05.

El cálculo de las estadísticas descriptivas para el segundo conjunto de datos se presenta en la tabla 3. En esta tabla se comprueba que con el segundo conjunto de datos sucede lo mismo que con el primero (mayores promedios se asocian a mayores notas en MPP), siendo el caso más representativo de esta tendencia el promedio en las ciencias.

Tabla 3. Estadísticas descriptivas para variables cuantitativas. Segundo conjunto

Calificación		Promedio ciencias	Promedio dibujo	Promedio inglés	Promedio técnicas	Promedio ciencias sociales
2	Media	2,9	4,08	4,0	4,12	4,2
	Desviación Estándar	0,62	0,67	0,62	0,67	0,47
3	Media	3,13	4,14	4,27	4,26	4,39
	Desviación Estándar	0,58	0,66	0,51	0,51	0,19
4	Media	3,73	4,45	4,51	4,54	4,6
	Desviación Estándar	0,69	0,53	0,46	0,58	0,26
5	Media	4,29	4,71	4,6	4,77	4,64
	Desviación Estándar	0,56	0,42	0,44	0,37	0,25

Al aplicar la RLO con el segundo conjunto de datos los resultados fueron los siguientes: .

⁴ En la sexta columna de la tabla 2 se presenta el valor del odds ratio para cada variable independiente y en las columnas siguientes se presenta un intervalo de confianza para este parámetro. Sólo se interpretará este estadístico en el modelo definitivo que se haya encontrado.

Tabla 4. RLO aplicada al segundo conjunto de datos

Link Function: Logit							
Logistic Regression Table							
Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI	
						Lower	Upper
Const(1)	17,5664	3,18104	5,52	0,000			
Const(2)	19,0910	3,24706	5,88	0,000			
Const(3)	21,4896	3,34975	6,42	0,000			
ININF							
3	-0,632467	1,46122	-0,43	0,665	0,53	0,03	9,31
4	-0,275095	1,43677	-0,19	0,848	0,76	0,05	12,69
5	-0,273031	1,46939	-0,19	0,853	0,76	0,04	13,56
Prom Cien	-1,45183	0,278102	-5,22	0,000	0,23	0,14	0,40
Prom CienSoc	-1,56848	0,631466	-2,48	0,013	0,21	0,06	0,72
Prom Dib	-0,107195	0,317565	-0,34	0,736	0,90	0,48	1,67
Prom Ingles	-0,714304	0,348294	-2,05	0,040	0,49	0,25	0,97
Prom Tecni	-0,728966	0,314518	-2,32	0,020	0,48	0,26	0,89
Log-Likelihood = -148,616							
Test that all slopes are zero: G = 96,853, DF = 8, P-Value = 0,000							
Goodness-of-Fit Tests							
Method	Chi-Square	DF	P				
Pearson	346,732	412	0,991				
Deviance	291,687	412	1,000				

En este caso, la prueba del aporte conjunto y las de bondad de ajuste llevan a las mismas conclusiones que con el primer conjunto de datos. El análisis individual de las variables independientes a través del estadístico Z de Wald, muestra que en este caso tuvieron una relación significativa con la nota en MPP, el promedio del estudiante en las ciencias, el promedio en las ciencias sociales, el promedio en inglés y el promedio en las asignaturas técnicas.

Con ambos conjuntos de datos fusionados los resultados son:

Tabla 5. RLO aplicada a ambos conjuntos de datos.

Link Function: Logit							
Logistic Regression Table							
Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI	
						Lower	Upper
Const(1)	16,9651	2,46705	6,88	0,000			
Const(2)	18,1348	2,49937	7,26	0,000			
Const(3)	20,0000	2,55340	7,83	0,000			
ININF							
3	0,299110	1,12499	0,27	0,790	1,35	0,15	12,23
4	0,187978	1,10762	0,17	0,865	1,21	0,14	10,58
5	1,03310	1,12709	0,92	0,359	2,81	0,31	25,59
Prom Cien	-1,32678	0,221554	-5,99	0,000	0,27	0,17	0,41
Prom CienSoc	-0,792416	0,454551	-1,74	0,081	0,45	0,19	1,10
Prom Dib	-0,423882	0,224863	-1,89	0,059	0,65	0,42	1,02
Prom Tecni	-0,788311	0,233617	-3,37	0,001	0,45	0,29	0,72
Prom Ingles	-1,02949	0,271515	-3,79	0,000	0,36	0,21	0,61
Log-Likelihood = -273,659							
Test that all slopes are zero: G = 147,291, DF = 8, P-Value = 0,000							
Goodness-of-Fit Tests							
Method	Chi-Square	DF	P				
Pearson	714,488	781	0,957				
Deviance	528,956	781	1,000				

En esta situación resulta significativa la relación entre MPP y el promedio en las ciencias, el promedio en Inglés, el promedio en las asignaturas técnicas, e incluso podría decirse que con el promedio en Dibujo, pues el valor p está muy cercano al nivel de significancia.

Las variables regresoras que muestran relación significativa con la variable dependiente varían en los tres análisis realizados. El promedio en las ciencias sociales aparece relacionado con MPP en los análisis individuales de los dos grupos de datos, pero no en el análisis conjunto. Por otra parte, sucede que el promedio en Dibujo, en Inglés y en las asignaturas técnicas no manifiestan su relación con la variable dependiente en alguno de los dos análisis individuales. Las soluciones halladas no son estables, pues varían en dependencia de los datos, indicando que las particularidades de los dos cursos que representan estos datos influyen en las relaciones encontradas. Solamente la variable explicativa “promedio en las ciencias” coincide en los tres casos como significativa, demostrando la independencia de su efecto de un conjunto de datos en particular. Por ser ésta la variable que con más seguridad se puede afirmar que influye en los resultados de MPP, valorando que es un primer estudio que se realiza con fines exploratorios y con propósitos de ser continuado, se decide solamente incluir esta variable en el modelo.

4.1. ANÁLISIS Y EVALUACIÓN DEL MODELO

Una vez decidido que lo más adecuado es construir el modelo incluyendo como variable regresora solamente el “promedio en las ciencias”, se obtiene la salida del software sólo con esta variable como predictora y los 274 estudiantes (ambos conjuntos de datos agrupados). Los resultados son:

Tabla 6. RLO con el promedio en ciencias como regresora (ambos conjuntos de datos agrupados).

Link Function: Logit								
Logistic Regression Table								
Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI		
						Lower	Upper	
Const(1)	5,63773	0,677181	8,33	0,000				
Const(2)	6,63092	0,710571	9,33	0,000				
Const(3)	8,22901	0,770060	10,69	0,000				
Prom Cien	-1,69095	0,190936	-8,86	0,000	0,18	0,13	0,27	

Log-Likelihood = -299,752
 Test that all slopes are zero: G = 95,105, DF = 1, P-Value = 0,000

Goodness-of-Fit Tests			
Method	Chi-Square	DF	P
Pearson	43,6466	44	0,487
Deviance	46,9835	44	0,351

Los valores p de las pruebas de falta de ajuste son mayores que 0,05, lo que permite concluir que con la ecuación se alcanza un buen ajuste. El coeficiente negativo del promedio en ciencias hace que el *odds ratio* sea menor que uno, indicando que incrementos en este promedio aumentan la probabilidad de obtener resultados más altos y, por lo

tanto, mejores en MPP. El hecho de que el intervalo de confianza del 95% de confianza para el *odds ratio* no incluya al uno permite asegurar, con una probabilidad menor que 0,05 de estar equivocado, que es significativa la disminución en la razón de probabilidad acumulada causada por el aumento del promedio (Agresti, 1990). Específicamente, el valor de 0,18 del *odds ratio* indica que un incremento unitario en el promedio en ciencia disminuye hasta un 18% simultáneamente, la razón entre la probabilidad que el estudiante suspenda (obtenga 2) y la probabilidad que aprueba (obtenga al menos 3); la razón entre la probabilidad que el estudiante obtenga 3 o menos y la probabilidad que saque al menos 4, y la razón entre la probabilidad que saque 4 o menos y la probabilidad que saque 5.

El modelo obtenido es:

$$P(\text{calificación MPP} \leq \text{categoría } i / \text{valor del promedio}) = e^{\alpha_i - 1,69 \text{ prom}} / 1 + e^{\alpha_i - 1,69 \text{ prom}}$$

La primera categoría es 2 (mala calificación), la segunda es 3 (calificación regular) y la tercera es 4 (buena calificación). Los α_i serían:

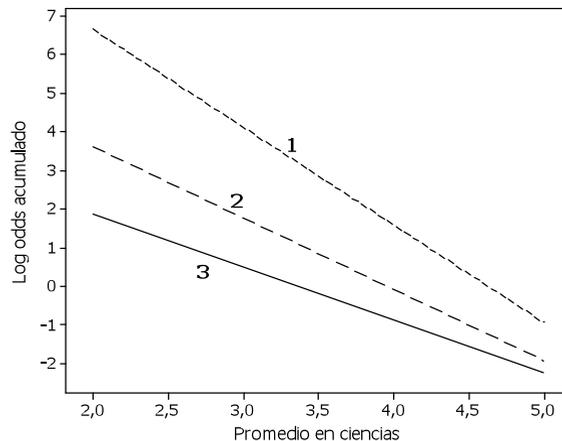
$$\alpha_1 = 5,64$$

$$\alpha_2 = 6,63$$

$$\alpha_3 = 8,23$$

Para comprobar que se cumple el requisito de proporcionalidad, lo cual significa, en este caso, que las variaciones en el promedio en las ciencias producen el mismo cambio en la razón de probabilidad acumulada de todas las categorías, se construye la Figura 1:

Figura 1. Relación entre el promedio en ciencias y el logaritmo del *odds* acumulado



La recta denotada con el uno muestra la relación entre el logaritmo del *odds* acumulado hasta el valor cuatro de la variable dependiente, y el promedio en las ciencias. La recta marcada con el dos representa la relación entre el logaritmo del *odds* acumulado

hasta el valor tres de la variable dependiente y el promedio en las ciencias. Por último, la recta que se señala con el tres muestra la relación entre el logaritmo del *odds* acumulado hasta la calificación de dos en MPP y el promedio en las ciencias. No se observa una falta de paralelismo notable entre las tres rectas, lo que indica que el efecto del promedio en las ciencias es el mismo sobre las razones de probabilidad acumulada de todas las categorías, pudiendo darse por cumplido el criterio de proporcionalidad.

Para evaluar finalmente la efectividad del modelo haciendo uso del Minitab, se sustituyó el valor del “promedio en las ciencias” de los 145 estudiantes de primer año del curso 2010-2011 y se obtuvo, antes de comenzar la impartición de la asignatura MPP, una predicción del rendimiento individual de cada alumno en esta materia durante el curso 2011-2012. Para realizar esta predicción se tiene en cuenta que como resultado del uso de la RLO lo que se logra es una estimación de la probabilidad que el estudiante obtenga cada una de las posibles calificaciones, es decir, 2, 3, 4 y 5. Atendiendo a esta estimación, la predicción se estructuró de la siguiente manera:

- No se entendió adecuado proponer como resultado de la predicción los eventos simples (2, 3, 4 y 5), pues aun cuando alguno de estos eventos en ciertos casos era más probable que los otros (tenía una probabilidad estimada mayor), sucedía que esta probabilidad era realmente baja (por ejemplo 0,35). Este inconveniente llevó a estructurar la predicción basándose, en ocasiones, en eventos compuestos (que contienen más de un punto muestral). De esta forma se pronosticó alguno de estos eventos:

- El estudiante reprueba la asignatura (obtiene 2).
- El estudiante aprueba la asignatura (obtiene al menos 3).
- El estudiante obtiene buena calificación en la asignatura (obtiene al menos 4).

- Se pronosticó para el estudiante alguno de estos tres eventos si la probabilidad predicha para alguno era superior a 0,7. Se utiliza el valor de 0,7 para desarrollar la predicción pues se considera insuficiente emplear, como se recomienda en algunos casos, 0,5 como probabilidad de referencia (Luan, 2002).

Vale destacar que con este método hay estudiantes a los que no se les logra pronosticar resultado, lo cual ocurre, bien porque al acumular los eventos simples no se logra que ninguno de los tres eventos mencionados tenga una probabilidad superior a 0,7, o bien porque la unión de eventos simples con la que esto se logra es la de 2 con 3, lo cual se juzga como un pronóstico poco útil desde el punto de vista práctico, atendiendo a que 2 implica estar suspenso y 3 aprobado. Las predicciones se compararon con los resultados reales de los alumnos, los cuales se obtuvieron al final del primer semestre del curso. El cálculo del porcentaje total de aciertos derivado de esta comparación constituye la base empleada para juzgar la efectividad del modelo. Este mismo criterio fue el empleado por Reyes *et al.*, (2007) en un trabajo previo, publicado en esta revista, donde los autores plantean que escogen como mejor solución aquella con la cual el porcentaje total de aciertos es máximo. Con los datos de los 145 estudiantes se confeccionó una tabla donde se incluyó: las probabilidades estimadas, el pronóstico realizado, el resultado real y si hubo o no acierto.

A manera de ejemplo, se expone en la tabla 7 el pronóstico realizado para cuatro de los 145 alumnos. Esta representa una muestra de la que fue confeccionada para todo el conjunto de estudiantes.

Tabla 7. Ejemplo de la predicción realizada en segundo año (curso 2011-2012)

Estudiante	Prob de 2	Prob de 3	Prob de 4	Prob de 5	Pronóstico	Nota real	Acierto
Estudiante X	0,08	0,11	0,34	0,47	Buena Calificación	4	Sí
Estudiante Y	0,24	0,22	0,35	0,19	Aprobar	3	Sí
Estudiante Z	0,78	0,13	0,07	0,02	Reprobar	3	No
Estudiante K	0,47	0,24	0,21	0,08	Sin pronóstico		

El estudiante denotado como X (se emplea esta denominación para mantener el anonimato de los estudiantes implicados) acumula una probabilidad de 0,81 entre los resultados 4 y 5; es por eso que se pronostica una buena calificación. Como finalmente obtiene 4, se registra como un acierto logrado en el pronóstico. Para el estudiante Y se pronostica que apruebe (obtenga al menos 3) con una probabilidad de 0,76; como realmente logra 3 en la asignatura, su predicción fue un acierto. Para el estudiante Z se pronostica que suspenda pues se estima una probabilidad de 0,78 de obtener calificación de “2” puntos, pero en este caso la predicción falló, pues este alumno logró 3 puntos como calificación. Con el estudiante K no fue posible encontrar una combinación de probabilidades estimadas que permitiera, según el método de predicción adoptado, pronosticar su resultado.

En la Tabla 8 se resumen los resultados del pronóstico ejecutado con los 145 educandos.

Tabla 8. Resumen del pronóstico de MPP (curso 2011-2012)

Total estudiantes	Total sin pronóstico	Total con pronóstico	
145	42	103	
		82	Total aciertos
		21	Total fallos

De los 103 estudiantes cuyo pronóstico fue realizado, hubo 82 aciertos y 21 errores, para un 80 % de aciertos y un 20 % de fallos. En otro estudio realizado en el ISPJAE que empleó redes neuronales para predecir la calificación de los alumnos en Inglés, se logró una efectividad del 83,4% (Alfonso, 2008). En nuestro mismo instituto, otra investigación donde se emplearon los árboles de clasificación para predecir el rendimiento académico de los alumnos de todas las carreras al finalizar el primer año, arrojó como resultado modelos con efectividad que varió entre 71 y 83% (Brito, 2008). En la Facultad de Matemática de la Universidad Central de las Villas se empleó un modelo híbrido basado en instancias para pronosticar si los alumnos tendrían o no éxito en su primer año, lográndose una efectividad de 73,5 %, al combinar los resultados conseguidos para las cinco carreras en donde se aplicó (King, 2008). Al comparar los resultados alcanzados en los estudios realizados en nuestro país con los del presente trabajo, se observa que la

efectividad del modelo logístico obtenido está dentro de los márgenes en que generalmente se ha encontrado este indicador, incluso superando en calidad a algunos de los modelos logrados. Vandamme *et al.* (2007) plantean que la calidad de los modelos de predicción académica es satisfactoria si la tasa de acierto está cercana o supera el 80%, valor que precisamente fue el obtenido para este caso de estudio. Teniendo en cuenta todas estas consideraciones, la precisión del modelo obtenido se juzga como adecuada.

Haciendo un análisis más detallado de la precisión del modelo, se pudo comprobar que de los 54 alumnos que se predijo suspenderían, 34 realmente suspendieron (63 %) y los otros 20 aprobaron (37 %). Por lo tanto la ecuación no tuvo una alta precisión al pronosticar los suspensos, pero este tipo de error es, de todos, el de menos consecuencias, ya que lo que sucede realmente es más positivo que lo pronosticado. De los 28 alumnos cuyo pronóstico fue “aprobar” (sin precisar con qué calificación), 27 aprobaron realmente, para una precisión del 97%. De los 21 estudiantes que se pronosticó obtendrían buena calificación, todos cumplieron la predicción, de manera que los errores más graves, que son aquellos en los que el pronóstico es más optimista que la realidad, es decir, pronosticar que el alumno va a aprobar y que realmente suspenda, o pronosticar que obtendrá buena calificación y que obtenga 3 o suspenda, presentan una menor incidencia en este caso. Este análisis más detallado confirma la adecuación del modelo, ya que su comportamiento más favorable resulta con los errores más graves.

Finalmente, de los 42 estudiantes a los que no se les predijo resultado, sólo 3 (7%) reprobaron y 39 aprobaron la materia (93%). De estos últimos, 23 obtuvieron 3 puntos y 16 obtuvieron 4. Por tanto, la evidencia empírica indica que los educandos que, según el método, quedan imprecisos para asignar a un resultado predictivo, son alumnos de bajo riesgo, ya que la mayoría aprueba finalmente la materia.

5. CONCLUSIONES Y DISCUSIÓN

De todas las variables independientes que se valoraron, la que demostró poseer una relación más estable con el resultado del alumno en MPP, fue el promedio del educando en las asignaturas de ciencia durante su primer año. De ahí que se construyó una ecuación para relacionar la calificación en MPP, como variable dependiente, con el promedio en las ciencias en primer año como variable independiente, para lo cual se emplearon los datos de 274 estudiantes pertenecientes a dos cursos académicos distintos. Se identificó que mayores valores de este promedio están asociados a mejores notas en MPP. Las estimaciones de probabilidad conseguidas con este modelo se usaron como base para el desarrollo de un método que permitió predecir el estado en la asignatura de los alumnos que la recibieron a inicios del curso 2011-2012. En esta predicción se logró un porcentaje de acierto general del 80%, el cual se juzga como adecuado. La precisión lograda individualmente para los diferentes estados que se predijeron también se valora como adecuada.

Es oportuno seguir perfeccionando el método de análisis propuesto en relación a varios aspectos. En primer lugar, con el método de pronóstico presentado, algunos alumnos quedan sin predicción de acuerdo a sus probabilidades estimadas. En este caso de estudio, el porcentaje de alumnos que perteneció a esta categoría fue de 29%. A pesar de que esto constituye una limitante del método, deben realizarse otros análisis, usando esta metodología para ver si se observa la misma tendencia presente en este estudio, según la

cual, la gran mayoría de los alumnos de este tipo finalmente aprueban la materia, y, de ser así, pudiera valorarse predecir, con baja probabilidad de error, como “aprobados” a los alumnos que tengan esta condición. Otro aspecto que puede ser valorado es la conveniencia de usar otras probabilidades de corte distintas al valor de 0,7 sugerido en este trabajo, para estudiar si se mejora la efectividad del modelo. También sería provechoso examinar la posibilidad de incluir nuevas variables independientes que reflejen otros aspectos distintos al rendimiento previo del educando. Por último, los autores entienden que, dado que el promedio en las ciencias demostró una relación estable y fuerte con las calificaciones en MPP, sería útil hacer un análisis más detallado donde las variables independientes sean las calificaciones en cada una de las materias de ciencia.

Es necesario tener en cuenta que en estos estudios de predicción diversos factores provocan la divergencia entre lo real y el pronóstico, como por ejemplo: los cambios en las características de aprendizaje de los estudiantes que cursan la asignatura de un año a otro, los cambios de profesores, las modificaciones en el contenido, secuencia de actividades y en el sistema de evaluación de la asignatura, entre otros. Si se presentan diferencias marcadas en estos factores al comparar la situación en la cual se construye el modelo con el escenario en el cual se predice, deben esperarse mayores divergencias.

A pesar de los cambios que se proyectan para perfeccionar el modelo encontrado, su calidad actual es adecuada, por lo que permite informar a los profesores satisfactoriamente y de forma proactiva sobre el resultado docente que se espera obtendrá cada alumno. Esta información facilitaría el desarrollo de acciones diferenciadas con los educandos, por ejemplo:

- Los alumnos que se prediga obtendrán buena calificación, pueden ser aprovechados por los profesores para que los apoyen en los laboratorios que se imparten como parte del contenido de la materia MPP. Los docentes también pueden dedicar a estos estudiantes sesiones especiales de preparación para estimularlos a que participen en exámenes de premio⁵ y para que presenten investigaciones relacionadas con la asignatura en la Jornada Científica Estudiantil.⁶ Además, estos educandos podrían ser preparados desde el mismo momento en que se imparte la materia para que se conviertan en alumnos ayudantes.⁷

- Con los alumnos con pronóstico de reprobar, se podría establecer un control del aprendizaje más frecuente y personalizado, por ejemplo, a través del establecimiento de horarios de consulta para que de forma más individualizada se chequee la realización de los ejercicios orientados y se discutan las dificultades con el contenido que puedan existir.

- Se puede lograr una organización más adecuada de los equipos de trabajo que tradicionalmente el profesor crea en la sala de clase para resolver los problemas en las actividades prácticas, de manera que en estos equipos haya un equilibrio en cuanto al rendimiento de cada uno de sus miembros.

⁵ Son exámenes de mayor complejidad que se realizan por cada asignatura al final del semestre y cuyos ganadores reciben una bonificación adicional para su promedio académico.

⁶ La Jornada Científica Estudiantil es un evento que se celebra anualmente y donde los alumnos de todos los años académicos presentan investigaciones de diversa índole relacionadas con su perfil profesional.

⁷ En la educación superior cubana se denomina así a los alumnos que bajo la guía y el control de un profesor, imparten las actividades prácticas de una determinada materia a alumnos de cursos inferiores.

• Por otra parte, al impartir la materia se realizan dos seminarios para lo cual los estudiantes se agrupan según su preferencia en colectivos de tres a cuatro miembros. La información sobre el rendimiento esperado pueden ayudar a los docentes a establecer el orden en que deben exponer los equipos en estos seminarios, pues parece suceder que si en principio se exponen los mejores trabajos (que deben ser los realizados por los estudiantes de mayor desempeño), esto motiva al resto a elevar la calidad de sus presentaciones.

En resumen, la utilización de herramientas matemáticas que apoyen la planificación metodológica en materias donde la tasa de reprobación no es despreciable constituye una importante vía para elevar las posibilidades de éxito de los alumnos. La metodología expuesta en este artículo muestra sus potencialidades en este sentido por lo que sería acertado extender su aplicación a otras materias.

REFERENCIAS BIBLIOGRÁFICAS

- Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.
- Alfonso, D. (2008). Descubrimiento de patrones y reglas en el aprendizaje del idioma inglés utilizando técnicas de minería de datos. Tesis de diploma, ISPJAE, La Habana.
- Brito, R. (2008). Minería de Datos aplicada a la Gestión Docente del Instituto Superior Politécnico “José Antonio Echeverría”. Tesis de maestría, ISPJAE, La Habana.
- Delavari, N. & Beikzadeh, M. (2008). Data Mining Application in Higher Learning Institutions. *Informatics in Education*, vol. 7, n. 1, 31-54.
- Horruitiner, P. (2007). El problema de la calidad, el acceso y la pertinencia. *Pedagogía Universitaria*, vol. 12, n. 4, 92-132.
- Hosmer, D., Hosmer, T., Le Cessie, S. & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, vol. 16, 965-980.
- Hosmer, D. & Lemeshow, S. (2000). *Applied Logistic Regression*. New York: John Wiley & Sons.
- King, O. (2008). Sistema basado en instancias para pronosticar el éxito de un estudiante a su ingreso en una carrera universitaria. Tesis de diploma, UCLV, Villa Clara.
- Llinás, H. (2006). Precisiones en la teoría de los modelos logísticos. *Revista Colombiana de Estadística*, vol. 29, 239-265.
- Luan, J. (2002). Data Mining and Knowledge Management in Higher Education: Potential Applications. Paper presented at The Annual Forum for the Association for Institutional Research, Toronto.
- McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society*, vol. 42, 109-142.
- Organización Internacional de Normalización. (2002). *IWA-2: Sistemas de gestión de la calidad para organizaciones educativas*. México: Organización Internacional de Normalización.
- Ponsot, E., Sinha, Surendra. & Goitía, A. (2009). Sobre la agrupación de niveles del factor explicativo en el modelo logit binario. *Revista Colombiana de Estadística*, vol. 32, 157-187.
- Reyes, J., Escobar, C., Duarte, J. & Ramírez, P. (2007). Una aplicación del modelo de regresión logística en la predicción del rendimiento estudiantil. *Estudios Pedagógicos*, vol. 33, n. 2, 101-120.
- Vandamme, J-P, Meskens, N. & Superby, J-F. (2007). Predicting Academic Performance by Data Mining Methods. *Education Economics*, vol. 15, no. 4, 405-419.
- Zúnica, L., Alcover, R. & Valiente, J. (2005). Relación entre el rendimiento de dos asignaturas de segundo curso y las asignaturas de primer curso en Ingenierías Técnicas de Informática de la UPV. Ponencia presentada en XI Jornadas de Enseñanza Universitaria de la Informática, Madrid.