Estudios Pedagógicos XXXIII, Nº 2: 101-120, 2007

INVESTIGACIONES

*UNA APLICACION DEL MODELO DE REGRESION LOGISTICA EN LA PREDICCION DEL RENDIMIENTO ESTUDIANTIL**

An application of the model logistic regression in the prediction of the student performance

Jimmy Reyes Rocabado, Carlos Escobar Flores, Juan Duarte Vargas, Pedro Ramírez Peradotto

Universidad de Antofagasta, Departamento de Matemáticas, Avenida Angamos 601, Antofagasta, Chile. jreyes@uantof.cl

Resumen

Este trabajo presenta una metodología para realizar predicciones del éxito en el primer semestre de los estudiantes de la carrera de Ingeniería Plan Común, en una cohorte estudiantil de primer año de la Universidad de Antofagasta.

Para realizar los análisis se consideraron tres criterios de exigencia para clasificar como exitoso a un estudiante en el primer semestre de su carrera. Aplicando un modelo de regresión logística, los resultados fueron comparados con los del método de análisis discriminante, analizando además su concordancia e índice de predictibilidad.

Palabras clave: rendimiento estudiantil, regresión logística, concordancia, índice de predictibilidad.

Abstract

This work presents a methodology to carry out predictions of the success in the first semester of the students of the Common Plan Engineering career, in a student cohort of first year of the University of Antofagasta.

To carry out the analyses were considered three criteria of demand for classify as successful to a student in the first semester of their career. Applying a model of logistic regression, the results were compared with the of the discriminant analysis method, analyzing besides their agreement and index of predactibility.

Key words: student performance, logistic regression, agreement, index of predactibility.

* Proyecto de investigación financiado por la Dirección de Investigaciones de la Universidad de Antofagasta.

INTRODUCCION

Los directivos superiores asociados a la actividad docente de instituciones de educación superior asumen que la mayoría de los estudiantes que ingresan a una carrera universitaria lo hacen con la intención de permanecer en ella hasta su graduación y les resulta disonante aceptar que esos estudiantes, en altos porcentajes, puedan no tener éxito en varias de las asignaturas que cursan en el primer semestre de ingreso, debiendo prolongar sus estudios hasta casi duplicar sus tiempos normales para titulación. A su vez, la mayoría de ellos asume que las altas tasas de reprobación estudiantil (no éxito) se deben principalmente a métodos de enseñanza inadecuados, instrumentos de medición de resultados del aprendizaje no isométricos respecto a los resultados del aprendizaje, contenidos complejos, por su nivel de abstracción, contenidos innecesarios (a su juicio) para el futuro ejercicio profesional y mal diseño de la estructura de prerrequisitos entre asignaturas (Ramírez *et al.* 2004).

Los encargados de realizar la actividad docente asumen que las altas tasas de reprobación se deben a fallas de formación de los estudiantes lograda en enseñanza media, carencia de responsabilidad y de perseverancia del estudiante, problemas de inteligencia y carencia de aptitudes verbales y/o matemáticas (Ramírez *et al.* 2004).

Ante la situación descrita en forma general se hace necesario definir las variables que puedan estar involucradas en este mal rendimiento estudiantil durante el semestre inicial de su carrera y mediante un modelo relacionar estas variables con el rendimiento de tal forma de poder predecirlo con alta precisión antes de que este ocurra, para poder tomar las medidas preventivas.

METODOLOGIA

Para obtener el modelo de predicción se consideraron las siguientes variables:

VARIABLE DEPENDIENTE

Rendimiento: Variable por la cual se considera exitoso un rendimiento sobre la base de los tres siguientes criterios.

Criterio 1: Aprobar tres o más asignaturas en el primer semestre.
Criterio 2: Aprobar cuatro o más asignaturas en el primer semestre.
Criterio 3: Aprobar todas las asignaturas en el primer semestre.

VARIABLES INDEPENDIENTES

1. *Expectativa (EX):* Juicio que hace una persona respecto a su rendimiento esperado (por ejemplo, aprobar una signatura), que haga posible un logro deseado (por ejemplo, ser promovido de curso).
2. *Valencia (VA):* Importancia que asigna una persona a un resultado específico.
3. *Instrumentalidad (INS):* Valoración de la relación entre el esfuerzo que se realiza y lo que se logra sobre la base de ese esfuerzo.

4. *Puntaje de notas de enseñanza media (PtNt):* Puntaje asignado al promedio de calificaciones obtenida por el estudiante en Enseñanza Media.
5. *Puntaje en la PSU de matemáticas (PtMt):* Puntaje estandarizado correspondiente a la parte matemática de la PSU.
6. *Puntaje en la PSU de lenguaje (PtLg):* Puntaje estandarizado correspondiente a la parte de lenguaje de la PSU.
7. *Puntaje en ciencias (PtCs):* Puntaje estandarizado correspondiente a la prueba de ciencias de la PSU.

Los datos de las tres primeras variables corresponden a los tres constructos considerados por la teoría de Vroom y fueron obtenidos mediante un instrumento denominado "Inventario PEI-EDU-UA 93/94" estandarizado para alumnos de primer año universitario en una investigación anterior (Ramírez *et al.* 2004), financiada por la Dirección de Investigación de la Universidad de Antofagasta.

Los datos para las otras variables fueron obtenidos de la base de datos confeccionada por la Dirección de Admisión y Registro Curricular de la Universidad de Antofagasta (DARC).

EL MODELO

Para realizar las predicciones se aplicó un modelo de regresión logística, caracterizado de la siguiente forma (Barón López, Téllez Montiel 2000):

Si tenemos una variable que describe una respuesta en forma dicotómica (Éxito o fracaso) y queremos estudiar el efecto que otras variables (independientes) tienen sobre ella, el modelo de regresión logística binaria puede ser de gran utilidad para lo siguiente:

- Estimar la probabilidad de que se presente el evento de interés (por ejemplo, tener éxito en el primer semestre), dado los valores de las variables independientes,
- Evaluar la influencia que cada variable independiente tiene sobre la respuesta en forma de OR (ODD RATIO). Una OR mayor que uno indica aumento en la probabilidad del evento y una OR menor que uno, implica una disminución.

El modelo de regresión logística puede escribirse como:

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1 + b_2 x_2 + \ldots\ldots + b_k x$$

donde p es la probabilidad de que ocurra el evento de interés (en nuestro caso tener éxito en el primer semestre). Dado el valor de las variables independientes, podemos calcular directamente la estimación de la probabilidad de que ocurra el evento de interés de la siguiente forma:

$$\hat{p} = \frac{e^{suma}}{1+e^{suma}}\text{; donde } \; suma = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \ldots\ldots + \hat{b}_k x_k$$

En este caso también se puede establecer la significación de los coeficientes del modelo usando el estadístico de Wald (Lehmann 1974) que es equivalente al del modelo de regresión lineal múltiple. Si una variable independiente resulta no significativa podemos considerar eliminarla del modelo.

RESULTADOS OBTENIDOS

Para realizar el análisis se consideraron los tres criterios, mencionados anteriormente, para asignar éxito en el primer semestre:

Criterio 1: Aprobar tres o más asignaturas en el primer semestre.
Criterio 2: Aprobar cuatro o más asignaturas en el primer semestre.
Criterio 3: Aprobar todas las asignaturas en el primer semestre.

CRITERIO 1

Al realizar el análisis con este criterio y todas las variables involucradas, utilizando Statgraphics Plus 5.1, se observó que las variables puntaje en notas (PtNt) y puntaje en matemáticas de la PSU (PtMt) fueron las más significativas en el modelo (p–valor menor que 0,01), todas las demás fueron eliminadas del análisis.

Variable dependiente: RENDIMIENTO
Factores: PtNt, PtMt

Tabla 1

Modelo de regresión estimado (máxima probabilidad)

Parámetros	Estimadores	Error Estándar Estimado	Razón de Probabilidad
CONSTANTE	-16,1518	2,36825	
Coef. para PtNt	0,00585919	0,00196831	1,00598
Coef. para PtMt	0,0223587	0,00376586	1,02261

Tabla 2

Análisis de varianza

Fuente	Suma Cuadrados	Grados libertad	p - valor
Modelo	65,7905	2	0,0000
Residuos	221,959	214	0,3401
Total (corr.)	287,749	216	

Porcentaje de varianza explicado por el modelo = 22,8638%
Porcentaje ajustado = 20,7787%

Tabla 3

Test de razón de probabilidad

Factores	Chi-Cuadrado	Grados de libertad	p -Valor
PtNt	9,81218	1	0,0017
PtMt	49,2124	1	0,0000

La Tabla 1 muestra el resultado de ajustar un modelo de regresión logístico para describir la relación entre RENDIMIENTO según criterio 1 y las dos variables independientes.

La ecuación del modelo ajustado es

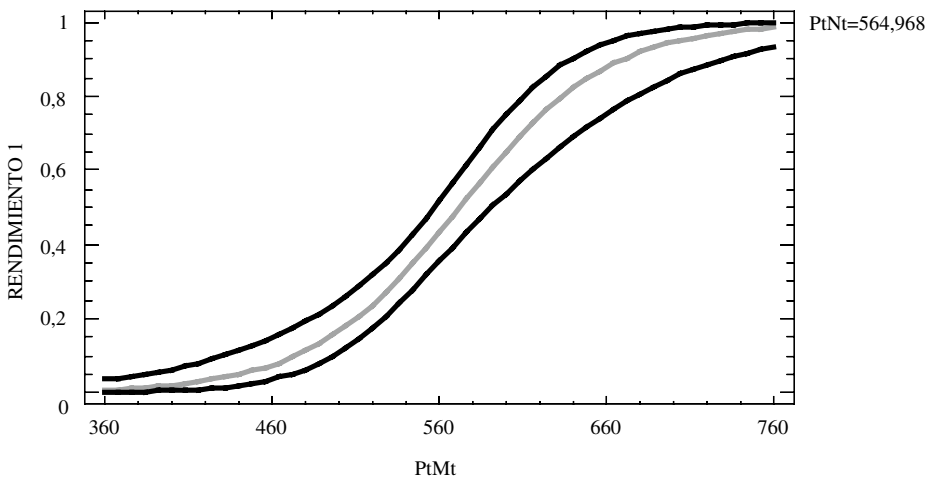
$$\hat{p} = P(RENDIMIENTO = 1) = \frac{e^{suma}}{1 + e^{suma}}$$

donde

suma = -16,1518 + 0,00595819*PtNt + 0,0223587*PtMt

Figura 1

Gráfico del modelo ajustado con 95,0% límites de confianza



Dado que el p-valor para el modelo en la tabla 2 de Análisis de la Varianza es inferior a 0,01, hay una relación estadísticamente significativa entre las variables al 99% de confianza.

Los resultados también muestran que el porcentaje de varianza de RENDIMIENTO explicado por el modelo es igual a 22,8638% sobre la base de las variables PtNt y PtMt, en este caso.

Para determinar el punto de corte (frontera) donde se clasifica a un estudiante como exitoso o no exitoso en el primer semestre, se calculan las siguientes probabilidades para diferentes puntos de corte usando el modelo de predicción:

- Probabilidad de predecir éxito dado que el estudiante tuvo éxito (Sensibilidad)
- Probabilidad de predecir no éxito dado que el estudiante no tuvo éxito (Especificidad)
- Probabilidad de acertar en la predicción para todos los estudiantes. (Total)

Los resultados se muestran en la siguiente tabla para diferentes puntos de corte (frontera):

Tabla 4

Determinación del punto de corte

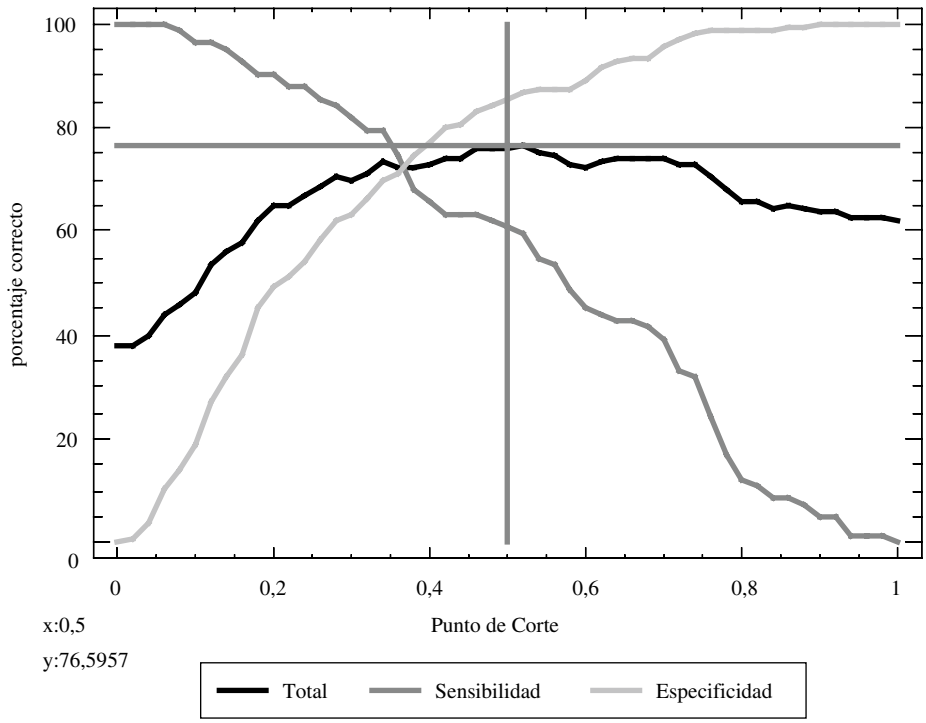
Frontera	Sensibilidad	Especificidad	Total
0	100	0	37,79
0,05	100	5,93	41,47
0,1	96,34	18,52	47,93
0,15	93,9	34,07	56,68
0,2	90,24	49,63	64,98
0,25	86,59	54,81	66,82
0,3	81,71	62,96	70,05
0,35	76,83	70,37	72,81
0,4	65,85	77,04	72,81
0,45	63,41	82,22	75,12
0,5	**60,98**	**85,19**	**76,04**
0,55	53,66	87,41	74,65
0,6	45,12	88,89	72,35
0,65	42,68	92,59	73,73
0,7	39,02	95,56	74,19
0,75	28,05	98,52	71,89
0,8	12,2	98,52	65,9
0,85	8,54	99,26	64,98
0,9	4,88	100	64,06
0,95	1,22	100	62,67
1	0	100	62,21

La tabla 4 muestra un resumen de la capacidad de predicción del modelo ajustado. En primer lugar, el modelo se utiliza para predecir la respuesta utilizando la información en cada fila del fichero de datos. Si el valor predicho es más grande que la frontera, se clasifica al estudiante como "exitoso". Si el valor predicho es inferior a o igual a la frontera, se clasifica al estudiante como "no exitoso". La tabla muestra el porcentaje de datos observados predichos correctamente a diferentes valores de la frontera. Por ejemplo, utilizando una frontera igual a 0,50, de todos los estudiantes exitosos el 60,98% se predijo correctamente, mientras que de todos los estudiantes no exitosos el 85,19% se predijo correctamente, con un 76,04% de predicciones correctas sobre el total de estudiantes.

El criterio utilizado para determinar el punto de corte (frontera) fue considerar el valor de frontera con el máximo porcentaje total de aciertos, además puede proporcionar un buen valor para utilizarlo en la predicción de valores adicionales.

Figura 2

Gráfico con Capacidad de predicción para Rendimiento 1



Con este criterio se realizaron las predicciones de rendimiento para los 217 estudiantes de la cohorte analizada y se comparó con los rendimientos reales obtenidos en el primer semestre de su carrera. Se obtuvo el siguiente cuadro:

Tabla 5

Distribución de estudiantes de primer año de la carrera de Ingeniería Plan Común ingresados el año 2004 a la Universidad de Antofagasta según rendimiento real y rendimiento predicho con modelo de regresión logística y Criterio 1

Rendimiento real	Predicción rendimiento		Total general
	Éxito	Fracaso	
Éxito	50	32	82
Fracaso	20	115	135
Total general	70	147	217

CRITERIO 2

Al realizar el análisis con este criterio y todas las variables involucradas, se observó que las variables Expectativa (EX), puntaje en notas (PtNt) y puntaje en matemáticas de la PSU (PtMt) fueron las más significativas en el modelo (p–valor menor que 0,01), todas las demás fueron eliminadas del análisis.

Variable dependiente: RENDIMIENTO
Factores: EX, PtNt, PtMt

Tabla 6

Modelo de regresión estimado (máxima probabilidad)

Parámetros	Estimadores	Error estándar estimado	Razón de probabilidad
CONSTANTE	-19,0875	3,03497	
Coef. para EX	1,63201	1,24212	5,11413
Coef. para PtNt	0,00331285	0,00231762	1,00332
Coef. para PtMt	0,0257899	0,00508723	1,02613

Tabla 7

Análisis de varianza

Fuente	Suma Cuadrados	Grados libertad	p - valor
Modelo	59,7373	2	0,0000
Residuos	159,091	213	0,9977
Total (corr.)	218,828	216	

Porcentaje de varianza explicado por el modelo = 27,2987%
Porcentaje ajustado = 23,6429%

Tabla 8

Test de razón de probabilidad

Factores	Chi-Cuadrado	Grados de libertad	p -Valor
EX	1,75848	1	0,1848
PtNt	2,07356	1	0,1499
PtMt	37,555	1	0,0000

La Tabla 6 muestra el resultado de ajustar un modelo de regresión logístico para describir la relación entre RENDIMIENTO con el criterio 2 y las tres variables independientes.

La ecuación del modelo ajustado es

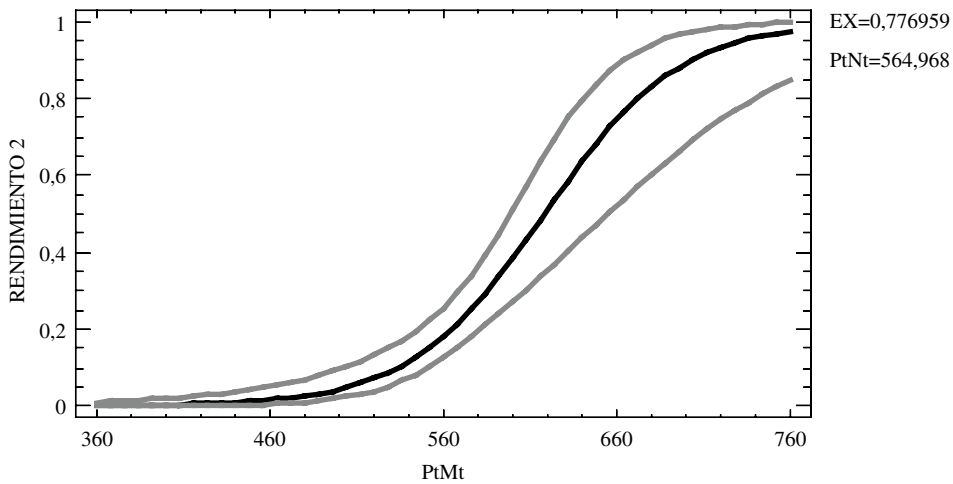
$$\widehat{p} = P(RENDIMIENTO = 1) = \frac{e^{suma}}{1 + e^{suma}}$$

donde

suma = -19,0875 + 1,63201*EX + 0,00331285*PtNt + 0,0257899*PtMt

Figura 3

Gráfico del Modelo Ajustado con 95,0% límites de confianza



Dado que el p-valor para el modelo en la tabla 7 de Análisis de la Varianza es inferior a 0,01, hay una relación estadísticamente significativa entre las variables al 99% de confianza.

Los resultados también muestran que el porcentaje de varianza de RENDIMIENTO explicado por el modelo es igual a 27,2987% sobre la base de las otras variables EX, PtNt y PtMt, en este caso.

Usando el mismo criterio anterior se puede obtener el punto de corte (Frontera) a partir de la siguiente tabla:

Tabla 9

Determinación del punto de corte

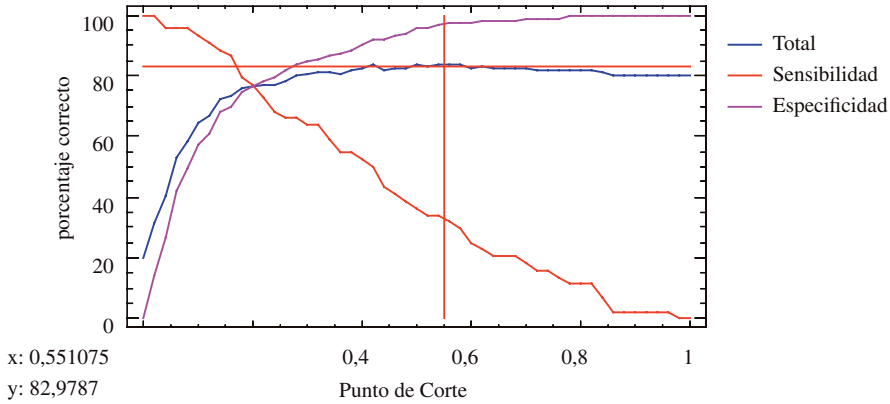
Frontera	Sensibilidad	Especificidad	Total
0	100	0	20,28
0,05	95,45	36,99	48,85
0,1	93,18	57,23	64,52
0,15	88,64	69,94	73,73
0,2	77,27	76,3	76,5
0,25	68,18	79,77	77,42
0,3	63,64	84,97	80,65
0,35	56,82	87,28	81,11
0,4	52,27	90,17	82,49
0,45	40,91	93,06	82,49
0,5	36,36	95,38	83,41
0,55	**34,09**	**96,53**	**83,87**
0,6	25	97,11	82,49
0,65	20,45	98,27	82,49
0,7	18,18	98,84	82,49
0,75	15,91	98,84	82,03
0,8	11,36	100	82,03
0,85	4,55	100	80,65
0,9	2,27	100	80,18
0,95	2,27	100	80,18
1	0	100	79,72

La tabla 9 muestra un resumen de la capacidad de predicción del modelo ajustado. En primer lugar, el modelo se utiliza para predecir la respuesta utilizando la información en cada fila del fichero de datos. Si el valor predicho es más grande que la frontera, se considera al estudiante como “exitoso”. Si el valor predicho es inferior a o igual a la frontera, se considera al estudiante como “no exitoso”. La tabla muestra el porcentaje de datos observados predichos correctamente a diferentes valores de frontera. Por ejemplo, utilizando una frontera igual a 0,55, de todos los estudiantes exitosos el 34,09% se predijo correctamente, mientras que de todos los estudiantes no exitosos el 96,53% se predijo correctamente, con un 83,87% de predicciones correctas sobre el total de estudiantes.

En este caso la capacidad de predicción del modelo se muestra en la figura 4:

Figura 4

Gráfico de capacidad de predicción para RENDIMIENTO 2



Con este criterio se realizaron las predicciones de rendimiento para los 217 estudiantes de la cohorte analizada y se comparó con los rendimientos reales obtenidos en el primer semestre de su carrera. Se obtuvo el siguiente cuadro:

Tabla 10

Distribución de estudiantes de primer año de la carrera de Ingeniería
Plan Común ingresados el año 2004 a la Universidad de Antofagasta según rendimiento real
y rendimiento predicho con modelo de regresión logística y Criterio 2

Rendimiento real	Predicción rendimiento		Total general
	Éxito	Fracaso	
Éxito	15	29	44
Fracaso	6	167	173
Total general	21	196	217

CRITERIO 3

Al realizar el análisis con este criterio y todas las variables involucradas, se observó que las variables puntaje en notas (PtNt), puntaje en matemáticas de la PSU (PtMt) y puntaje en ciencias (PtCs) fueron las más significativas en el modelo (p–valor menor que 0,01); todas las demás fueron eliminadas del análisis.

Variable dependiente: RENDIMIENTO
Factores: PtNt PtMt PtCs

Tabla 11

Modelo de regresión estimado (máxima probabilidad)

Parámetros	Estimadores	Error estándar estimado	Razón de probabilidad
CONSTANTE	-22,8795	4,27017	
Coef. para PtNt	0,0070403	0,00319071	1,00707
Coef. para PtMt	0,0203983	0,00679178	1,02061
Coef. para PtCs	0,00989672	0,00525345	1,00995

Tabla 12

Análisis de varianza

Fuente	Suma cuadrados	Grados libertad	p - valor
Modelo	41,7173	3	0,0000
Residuos	100,682	213	1,0000
Total (corr.)	142,4	216	

Porcentaje de varianza explicado por el modelo = 29,2959%
Porcentaje ajustado = 23,6779%

Tabla 13

Test de razón de probabilidad

Factores	Chi-Cuadrado	Grados de libertad	p -Valor
PtNt	5,19228	1	0,0227
PtMt	10,4475	1	0,0012
PtCs	3,89339	1	0,0485

La tabla 11 muestra el resultado de ajustar un modelo de regresión logístico para describir la relación entre RENDIMIENTO según criterio 3 y las tres variables independientes.

La ecuación del modelo ajustado es

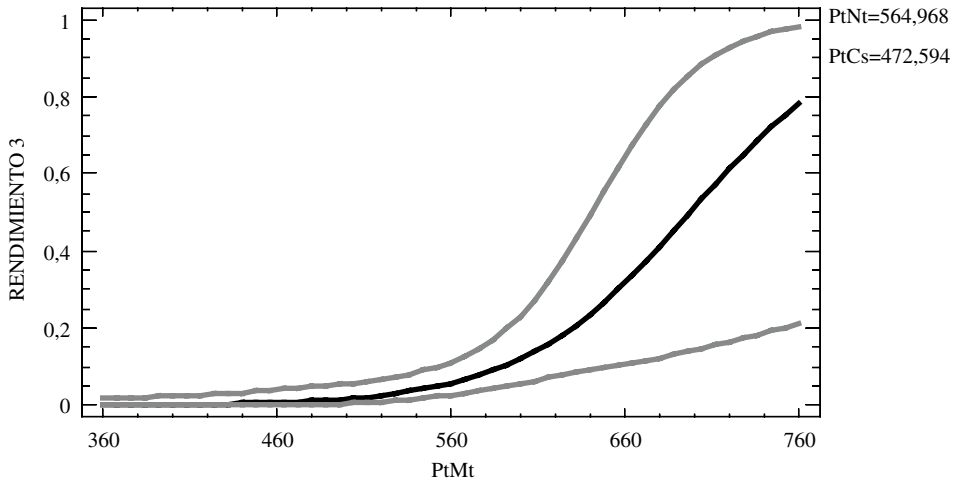
$$\hat{p} = P(RENDIMIENTO = 1) = \frac{e^{suma}}{1 + e^{suma}}$$

donde

suma = -22,8795 + 0,0070403*PtNt + 0,0203983*PtMt + 0,00989672*PtCs

Figura 5

Gráfico del Modelo Ajustado con 95,0% límites de confianza



Dado que el p-valor para el modelo en la tabla 12 de Análisis de la Varianza es inferior a 0,01 hay una relación estadísticamente significativa entre las variables al 99% de confianza.

La ventana también muestra que el porcentaje de varianza de RENDIMIENTO explicado por el modelo es igual a 29,2959% sobre la base de las variables PtNt, PtMt y PtCs, en este caso.

Usando el mismo criterio anterior se puede obtener el punto de corte (Frontera) a partir de la tabla 14 (ver página siguiente).

La tabla 14 muestra un resumen de la capacidad de predicción del modelo ajustado. En primer lugar, el modelo se utiliza para predecir la respuesta utilizando la información en cada fila del fichero de datos. Si el valor predicho es más grande que la frontera, se considera al estudiante como "exitoso". Si el valor predicho es inferior a o igual a la frontera, se considera al estudiante como "no exitoso". La tabla muestra el porcentaje de datos observados predichos correctamente a diferentes valores de frontera. Por ejemplo, utilizando una frontera igual a 0,50, de todos los estudiantes exitosos el 27,27% se predijo correctamente, mientras que de todos los estudiantes no exitosos el 98,97%, se predijo correctamente, con un 91,71% de predicciones correctas sobre el total de estudiantes.

Tabla 14

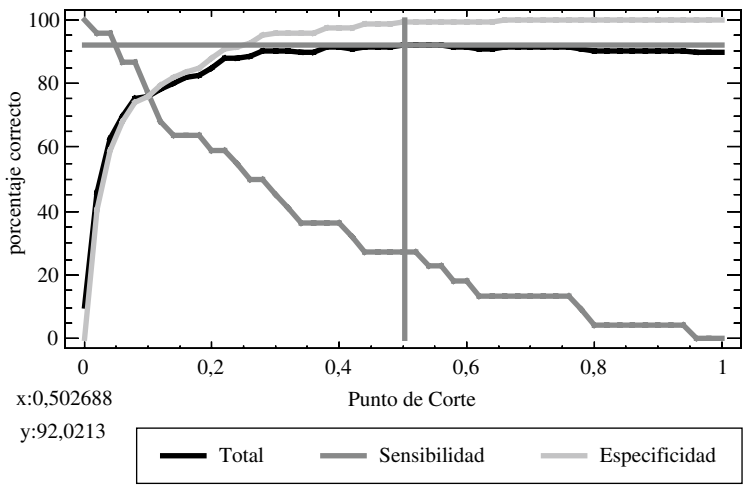
Determinación del punto de corte

Frontera	Sensibilidad	Especificidad	Total
0	100	0	10,14
0,05	95,45	62,05	65,44
0,1	77,27	75,9	76,04
0,15	63,64	82,56	80,65
0,2	59,09	87,69	84,79
0,25	50	91,79	87,56
0,3	45,45	95,38	90,32
0,35	36,36	95,9	89,86
0,4	36,36	97,44	91,24
0,45	27,27	98,46	91,24
0,5	**27,27**	**98,97**	**91,71**
0,55	22,73	99,49	91,71
0,6	18,18	99,49	91,24
0,65	13,64	100	91,24
0,7	13,64	100	91,24
0,75	13,64	100	91,24
0,8	4,55	100	90,32
0,85	4,55	100	90,32
0,9	4,55	100	90,32
0,95	0	100	89,86
1	0	100	89,86

En este caso la capacidad de predicción del modelo se muestra en la figura 6:

Figura 6

Gráfico de capacidad de predicción para RENDIMIENTO 3



Con este criterio se realizaron las predicciones de rendimiento para los 217 estudiantes de la cohorte analizada y se comparó con los rendimientos reales obtenidos en el primer semestre de su carrera. Se obtuvo la siguiente tabla:

Tabla 15

Distribución de estudiantes de primer año de la carrera de Ingeniería Plan Común ingresados el año 2004 a la Universidad de Antofagasta según rendimiento real y rendimiento predicho con modelo de regresión logística y criterio 3

Rendimiento real	Predicción rendimiento		Total general
	Éxito	Fracaso	
Éxito	6	16	22
Fracaso	2	193	195
Total general	8	209	217

RESULTADOS OBTENIDOS CON EL METODO DE ANALISIS DISCRIMINANTE

Utilizando el método de análisis discriminante (Morrison 1976) para los tres criterios se obtuvieron las siguientes tablas de clasificación:

Tabla 16

Distribución de estudiantes de primer año de la carrera de Ingeniería Plan Común ingresados el año 2004 a la Universidad de Antofagasta según rendimiento real y rendimiento predicho con método de análisis discriminante y criterio 1

Rendimiento real	Predicción rendimiento		Total general
	Éxito	Fracaso	
Éxito	57	25	82
Fracaso	33	102	135
Total general	90	127	217

Tabla 17

Distribución de estudiantes de primer año de la carrera de Ingeniería Plan Común ingresados el año 2004 a la Universidad de Antofagasta según rendimiento real y rendimiento predicho con método de análisis discriminante y criterio 2

Rendimiento real	Predicción rendimiento		Total general
	Éxito	Fracaso	
Éxito	33	11	44
Fracaso	42	131	173
Total general	75	142	217

Tabla 18

Distribución de estudiantes de primer año de la carrera de Ingeniería Plan Común ingresados el año 2004 a la Universidad de Antofagasta según rendimiento real y rendimiento predicho con método de análisis discriminante y criterio 3

Rendimiento real	Predicción rendimiento		Total general
	Éxito	Fracaso	
Éxito	16	6	22
Fracaso	45	150	195
Total general	61	156	217

VALIDACION DEL DIAGNOSTICO

Una forma de validar los resultados obtenidos con los modelos de predicción es usando el estadístico de Mc Nemar (Lehmann 1974) para los tres criterios. Para esto consideramos lo siguiente:

f_1= Número de estudiantes que tuvieron éxito en el primer semestre según criterio
f_2= Número de estudiantes considerados exitosos por el modelo de predicción

En este caso se puede estructurar la siguiente tabla:

Predicción Rendimiento	Rendimiento Real		Total general
	Fracaso	Éxito	
Fracaso	A	B	$n - f_2$
Éxito	C	D	f_2
Total general	$n - f_1$	f_1	n

Luego definiendo: p_1 = Proporción de éxitos según criterio
p_2 = Proporción de éxitos según el modelo de predicción

se tiene que $\hat{p}_1 = \frac{f_1}{n}$ y $\hat{p}_2 = \frac{f_2}{n}$

Para verificar si el modelo de predicción es válido se debe aceptar la hipótesis nula H_0: $p_1 = p_2$ v/s H_A: $p_1 \neq p_2$ a un nivel de significación α, para lo cual se usa el estadístico de Mc Nemar definido por:

$$J_0 = \frac{(B-C)^2}{B+C} \sim \chi^2(1) \text{ o } Z_0 = \frac{B-C}{\sqrt{B+C}} \sim N(0,1) \text{ si } B+C \geq 10$$

El modelo de predicción es válido con un nivel de significación α si $J_0 < \chi^2_{1-\alpha}(1)$ o si $-z_{1-\frac{\alpha}{2}} \leq Z_0 \leq z_{1-\frac{\alpha}{2}}$

Tabla 19

Resultados de validación obtenidos con el método de Regresión Logística

Criterio	J_0	Punto Crítico (α = 0,05)	Conclusión
1	2,77	3,85	Resultados Válidos
2	15,11	3,85	Resultados no válidos
3	10,89	3,85	Resultados no válidos

Tabla 20

Resultados de validación obtenidos con el método de Análisis Discriminante

Criterio	J_0	Punto Crítico (α = 0,05)	Conclusión
1	1,10	3,85	Resultados Válidos
2	18,13	3,85	Resultados no válidos
3	29,82	3,85	Resultados no válidos

INDICE KAPPA

Para establecer la concordancia de los resultados de rendimientos estimados con el modelo de predicción y los resultados reales se usa el Índice Kappa (Landis, Koch 1977), el cual se puede definir de la siguiente forma:

Dada una tabla estructurada de la siguiente forma:

Predicción rendimiento	Rendimiento real		Total general
	Fracaso	Éxito	
Fracaso	A	B	N_1
Éxito	C	D	N_2
Total general	M_1	M_2	N

El índice Kappa se define en este caso por:

$$\kappa = \frac{N(A+D) - (N_1M_1 + N_2M_2)}{N^2 - (N_1M_1 + N_2M_2)}$$

Tabla 21

Resultados obtenidos de la concordancia mediante el modelo de regresión logística comparados con el método de análisis discriminante

	Regresión Logística		Análisis Discriminante	
Criterio	Indice Kappa	Concordancia	Indice Kappa	Concordancia
1	0,48	Moderada	0,44	Moderada
2	0,38	Débil	0,40	Débil
3	0,37	Débil	0,27	Débil

INDICE DE PREDICTIBILIDAD

Definamos los siguientes eventos:
E = {Éxito en el primer semestre}
E^C = {Fracaso en el primer semestre}
D^+ = {Predicción de rendimiento positivo (éxito)}
D^- = {Predicción de rendimiento negativo (fracaso)}

Se define el **índice de predictibilidad de verdaderos positivos** como la probabilidad de tener éxito en el primer semestre, dado que su predicción fue de éxito y se denota por $P(E/D^+)$

Se define el **índice de predictibilidad de verdaderos negativos** como la probabilidad de fracasar en el primer semestre, dado que su predicción fue de fracaso y se denota por $P(E^C/D^-)$

Tabla 22

Resultados obtenidos del índice de predictibilidad mediante el modelo de regresión logística comparados con el método de análisis discriminante

	Regresión Logística		Análisis Discriminante	
Criterio	$P(E/D^+)$	$P(E^C/D^-)$	$P(E/D^+)$	$P(E^C/D^-)$
1	0,7143	0,7823	0,6333	0,8031
2	0,7143	0,8520	0,44	0,9225
3	0,75	0,9234	0,2623	0,9615

CONCLUSIONES

- El modelo de regresión logística es un buen procedimiento para predecir el éxito en el primer semestre si se toma un criterio no tan exigente para considerar el "éxito" (criterio 1), de esta forma la prueba de validación del método no rechaza la hipótesis de igualdad de la probabilidad de éxito estimada con la verdadera probabilidad de éxito con un nivel de significación no mayor a 0,05, sin embargo con los otros dos criterios esta hipótesis es rechazada. Análogamente se concluye si consideramos el método de análisis discriminante.
- A medida que el criterio para considerar exitoso a un estudiante es más exigente se puede observar que la sensibilidad va disminuyendo y la especificidad va aumentando, en la determinación del punto de corte.
- De los estudiantes que fueron diagnosticados como exitosos en el primer semestre de su carrera, usando el criterio 1 con el modelo de regresión logística, el 71,4% realmente tuvo éxito y este porcentaje se mantiene si usamos el criterio 2 y aumenta a 75% si usamos el criterio 3.
- De los estudiantes que fueron diagnosticados como no exitosos en el primer semestre de su carrera, usando el criterio 1 con el modelo de regresión logística, el 78,23% realmente no tuvo éxito y este porcentaje aumenta a 85,2% si usamos el criterio 2 y aumenta a 92,34% si usamos el criterio 3.
- De los estudiantes que fueron diagnosticados como exitosos en el primer semestre de su carrera, usando el criterio 1 con el método de análisis discriminante, el 63,3% realmente tuvo éxito y este porcentaje baja a 44% si usamos el criterio 2 y baja a 26,23% si usamos el criterio 3.
- De los estudiantes que fueron diagnosticados como no exitosos en el primer semestre de su carrera, usando el criterio 1 con el método de análisis discriminante, el 80,31% realmente no tuvo éxito y este porcentaje aumenta a 92,25% si usamos el criterio 2 y aumenta a 96,15% si usamos el criterio 3
- La concordancia, utilizando ambos procedimientos de predicción (regresión logística y análisis discriminante), resulta moderada si se toma un criterio no tan exigente para considerar el "éxito" (criterio 1).

REFERENCIAS

Dobson, A. (2002). *An Introduction to Generalized Linear Models.* Editorial Chapman y Hall.

Barón López, F. J., F. Téllez Montiel (2000). *Apuntes de Bioestadística:* 52-57.

Duarte, J., C. Escobar; J. Reyes, (1997). Un método de estimación de parámetros para modelar fenómenos biológicos. Anales VII Congreso de Matemática Capricornio COMCA 97, U. Católica del Norte, Antofagasta, Chile.

Duarte, J., C. Escobar; J. Reyes (1998). El problema de estimación de parámetros en modelos con error en las variables. Anales VIII Congreso de Matemática Capricornio COMCA 98, U. de Tarapacá, Arica, Chile.

Ramírez, P. *et al.* (2004). Motivación y rendimiento académico en cuatro carreras del área biológica de la U. de Antofagasta, cohortes 2004. Seminario de título. Facultad de Educación y Ciencias Humanas, U. de Antofagasta, Antofagasta, Chile.

Lehmann, E. L. (1974). *Nonparametrics-Statistical Methods Based on Ranks.* John Wiley & Sons. New York.

Landis J. R., G. G. Koch (1977). The measurement of observer agreement for categorical data. *Biometrics* 33: 159-174.

Morrison, D. F. (1976). *Multivariate Statistical Methods.* 2ª edición. New York: McGraw-Hill.

Duarte, J., C. Escobar, J. Reyes (1987). Efectos del cambio de ponderadores de los requisitos de ingreso en las carreras de la Universidad de Antofagasta, año 1986. Resúmenes del IX Encuentro Nacional de Investigadores en Educación. Pontificia Universidad Católica de Chile.

Duarte, J., C. Escobar, J. Reyes (1988). Dos métodos multivariados: Una aplicación en el ambiente educacional. Resúmenes de la Séptima Jornada de Matemáticas. Universidad. Católica del Norte, Antofagasta, Chile.