Uso del análisis multivariado para la clasificación y caracterización de ingresantes universitarios según su formación matemática previa

*Olga Ávila *, Eleonora Cerati *, Roberto Macías **, Claudia Redolatti *, Ingrid Schwer *, María Laura Taverna**

Resumen.
La enseñanza en los primeros cursos de la Universidad actualmente se encuentra afectada por la heterogeneidad del nivel de conocimientos alcanzado por los alumnos en los estudios previos y las carencias de los mismos que se aprecian en materias concretas como matemática. Es importante detectar los aspectos que pueden influir en el desempeño de los ingresantes para mejorar el rendimiento académico y la retención de los estudiantes evitando alumnos con alto riesgo de fracaso. El objetivo de este trabajo es describir la utilización de las técnicas multivariadas de análisis discriminante y regresión logística para identificar un conjunto de variables que permitan distinguir entre dos grupos: alumnos de la primer materia en Matemática, Matemática A, de las carreras que se dictan en la Facultad de Ingeniería Química de la Universidad Nacional del Litoral, y alumnos que deben realizar un curso tutorial como curso de apoyo para lograr una mejor formación en matemática.

Introducción
A partir del año 2000 se realiza un seguimiento estadístico para evaluar distintos aspectos del rendimiento en matemática de los alumnos ingresantes de las diferentes carreras de la Facultad de Ingeniería Química (FIQ) dependiente de la Universidad Nacional del Litoral (UNL), con sede en la ciudad de Santa Fe. Se dictan, entre otras, las carreras de Ingeniería Química (IQ), Licenciatura en Química (LQ), Ingeniería Industrial (II), Analista Industrial (AI), Ingeniería en Alimentos (IA), Químico Analista (QA) y Profesorado en Química (PQ). Estas carreras cuentan con un núcleo de materias básicas del cual forma parte el primer curso de Matemática, llamado Matemática A. Esta materia abarca contenidos de cálculo y de álgebra lineal.
A partir del año 2001 es un requisito obligatorio la aprobación del Curso de Articulación Disciplinar en Matemática para el cursado de Matemática A. Los ingresantes que no lo aprueban deben asistir a los Cursos de Articulación Disciplinar tutoriales durante el primer cuatrimestre de la carrera. De esta forma los alumnos ingresantes quedan separados en dos grupos: aquellos que deben realizar el Curso Tutorial de Matemática ya que necesitan de una mayor formación previa en los contenidos del nivel Polimodal y los que se considera que tienen la formación suficiente para realizar Matemática A.
Estudios realizados en trabajos anteriores con alumnos ingresantes en las carreras de

la Facultad, mostraron que la aprobación del Curso de Articulación en el área matemática es un factor importante y decisivo para un buen desempeño de los alumnos en el primer curso de Matemática [1]. Entonces surge la necesidad de analizar en mayor detalle distintos aspectos que diferencian a los alumnos que aprobaron este curso de aquellos que deben realizar el Curso Tutorial de Matemática para reforzar contenidos del polimodal.
El objetivo de este trabajo es describir resultados obtenidos de la utilización de las técnicas multivariadas de análisis discriminante y de regresión logística, para distinguir entre estos dos grupos. Ambos métodos permiten identificar, de un grupo numeroso, un conjunto de variables independientes como las variables mejores predictoras.
Este tipo de estudio puede utilizarse para hacer un diagnóstico sobre la formación actual de cada alumno, evaluando si es apropiada para un buen desempeño en el primer curso de matemática. Un objetivo es identificar alumnos que pueden tener alguna deficiencia aún teniendo aprobado el curso de Articulación, a los que se les prestará especial atención para evitar situaciones de fracaso.

Datos
Para la obtención de los datos se utilizaron los formularios SUR (Sistema Único de Registro), los resultados obtenidos del Curso de Articulación Disciplinar en Matemática, y una encuesta similar a la realizada en el año 2000 donde se recogen datos relacionados con el curso de articulación disciplinar y se realizan preguntas referidas al estudio de recta, parábola y trigonometría. Las respuestas a estas preguntas permiten reflejar aspectos de la formación obtenida en el nivel polimodal en el área matemática.

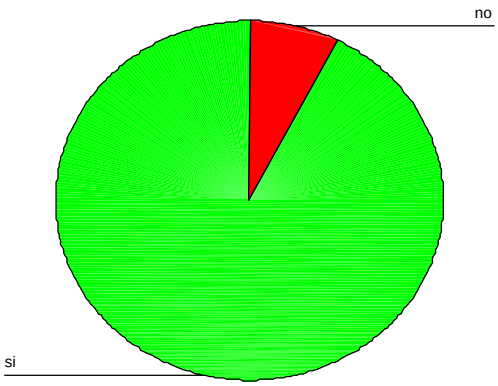
Variables
Inicialmente se consideran las siguientes variables:
a) Cursó Matemática en el último año del nivel polimodal (variable dicotómica).
b) Recuerda haber estudiado recta (variable dicotómica).
c) Recuerda haber estudiado parábola (variable dicotómica).
d) Recuerda haber estudiado trigonometría (variable dicotómica).
e) Situación previa: cuenta con estudios universitarios o terciarios previos (variable dicotómica).
f) Cantidad de conocimientos nuevos que proporciona el Curso de Articulación: se han considerado tres modalidades ninguno, pocos y muchos.
g) Dependencia del establecimiento donde cursó el polimodal (pública o privada).
h) Orientación del polimodal .
i) Nota del examen del Curso de Articulación en Matemática (variable continua).
j) Situación laboral: se han considerado las modalidades no trabaja, trabaja menos de 20 horas, entre 20 y 30 horas, y más de 30 horas.

Salvo la variable i), las restantes se obtienen de la encuesta mencionada.
Si bien en la encuesta la variable Orientación del polimodal (h) tiene varias opciones, a los efectos de este análisis fue reformulada como variable dicotómica considerando técnicos y no técnicos. Para la variable situación laboral también fue reformulada como trabaja y no trabaja.

Análisis Descriptivo de las variables que definen los modelos
El Software utilizado para el procesamiento de los datos fue el SPSS versión 9.0.
Se presentan a continuación las tablas de frecuencias asociadas a las variables que resultaron posteriormente significativas en los modelos estadísticos considerados en el presente trabajo.

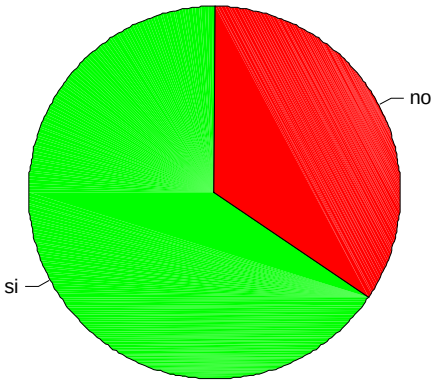
Recuerda haber estudiado parábola

	Frecuencia	Porcentaje
no	15	7,6
si	183	92,4
Total	198	100,0



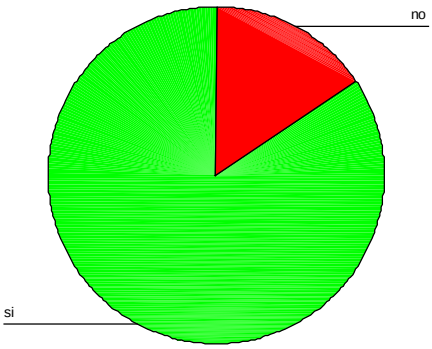
Tuvo Matemática el último año

	Frecuencia	Porcentaje
no	68	34,3
si	130	65,7
Total	198	100,0



Recuerda haber estudiado Trigonometría

	Frecuencia	Porcentaje
no	31	15,7
si	167	84,3
Total	198	100,0



Nota del Curso de Articulación de Apoyo

Nota	Frecuencia	Porcentaje
1,00	2	1,0
2,00	10	5,1
3,00	7	3,5
4,00	34	17,2
5,00	13	6,6
6,00	24	12,1
7,00	19	9,6
8,00	19	9,6
9,00	22	11,1
10,00	48	24,2
Total	198	100,0

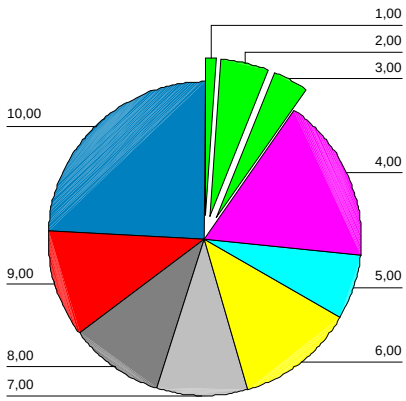


Tabla 1: Frecuencias de las variables seleccionadas para la construcción de los modelos discriminante y regresión logística

Análisis Discriminante

La técnica multivariada de análisis discriminante permite relacionar diferentes aspectos relevantes para discriminar en dos grupos para caracterizar la formación de los ingresantes en la Universidad. Es necesario tener grupos excluyentes y diferenciados. En este caso son dos: aquellos que están en condiciones de cursar Matemática A y quienes deben asistir al Curso de Apoyo.

Para utilizar esta técnica se necesita la información relacionada con dos aspectos esenciales: un conjunto de casos cuyo grupo de pertenencia está identificado y un conjunto de variables que identifiquen las características que mejor miden las diferencias entre esos grupos.

El análisis discriminante [3,4,5] permite construir una función que es una combinación lineal de variables que evaluada en cada caso proporciona un puntaje que permite asignarlo al grupo de pertenencia. Una vez que esta función discrimina correctamente los casos conocidos puede utilizarse para clasificar nuevos casos cuyo grupo de pertenencia se desconoce.

La técnica permite trabajar con grupos con una gran diferencia en cantidad de datos, utilizando probabilidades proporcionales al tamaño de los grupos, situación a tenerse en cuenta en el presente trabajo ya que se tienen solamente 19 alumnos que asistieron al Curso de Apoyo y 179 alumnos que estuvieron en condiciones de cursar Matemática A.

Una etapa importante en la técnica de análisis discriminante para construir un modelo adecuado es poder seleccionar correctamente un subconjunto de variables capaces de

discriminar entre los dos grupos identificados. Para tal fin inicialmente se realiza un test de igualdad de medias entre grupos a todas las variables disponibles. Las variables ¨Recuerda si estudió parábola¨ y ¨ Nota del examen del Curso de Articulación en Matemática¨ son significativas al nivel menor del 1 %, mientras que ¨Recuerda si estudió trigonometría¨ y ¨Tuvo Matemática en el último año¨ mostraron valores *p* de significación próximos al 5 % y las variables restantes no se consideran en la construcción del modelo por ser significativas según los niveles trabajados.
A las cuatro variables seleccionadas para la construcción del modelo se les realizó un test de hipótesis, resultando todas significativas para formar parte del mismo (tabla 2).

Variable	**Lambda de Wilks**	**F**	**Gl 1**	**Gl 2**	**Significación**
Estudió Parábola	0.948	11,022	1	196	0.001
Estudió Trigonometría	0.964	7.336	1	196	0.007
Nota Curso	0.676	93.95	1	196	0.000
Mat. Ultimo año	0.96	8.13	1	196	0.005

Tabla 2: Significación de las variables en el modelo

Existen otros estadísticos propios de la técnica, como el coeficiente de correlación canónica y Lambda de Wilks, que permiten elegir el modelo más apropiado de todos los modelos que pueden construirse a partir de las cuatro variables previamente seleccionadas. Ambos coeficientes miden las diferencias entre los grupos debidas a las funciones discriminantes. El estadístico Lambda de Wilks expresa la proporción de variabilidad total no debida a las diferencias entre los grupos (variabilidad dentro de los grupos); permite además contrastar la hipótesis nula de que las medias multivariantes de los grupos (centroides) son iguales. Al realizar esta prueba para nuestro caso se rechaza la hipótesis nula a una significación mayor al 1 %, lo cual indica que el modelo propuesto es apropiado para discriminar entre los grupos.

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	0,64	83,98	4	0,00

Tabla 3: Significación del estadístico Lambda de Wilks

El coeficiente de correlación canónica mide para la función discriminante, el grado en que difieren las medias de dicha función en los distintos grupos. Un valor alto del coeficiente indica una fuerte relación entre el grupo de pertenencia y los valores de la función discriminante. En este caso, resulta un valor moderado para este coeficiente.

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	0,54	100,0	100,0	0,59

Tabla 4: Correlación canónica

Función discriminante

La Tabla 5 detalla los coeficientes estandarizados de la función obtenida con las variables que finalmente resultaron significativas. El valor absoluto de estos coeficientes indica la importancia relativa de la variable correspondiente en el cálculo de la función discriminante.

	Coeficientes
Nota Curso Articulación	0,90
Estudió Parábola	0,29
Estudió Trigonometría	0,11
Matemática Ultimo	-0,09

Tabla 5: Coeficientes estandarizados de las funciones discriminantes canónicas

La función queda definida de la siguiente manera:

D_i = 0,294 estudió Parábola +0,116 Estudió Trigonometría + 0,904 Nota Apoyo – 0,089 Mat. Ultimo año

Esta función evaluada en cada caso arroja un puntaje D_i, que permite establecer el grupo de pertenencia del alumno.

Para poder interpretar la Tabla 6 se debe recordar que se desea discriminar el grupo formado por alumnos que deben realizar el Curso Tutorial de Matemática es decir que no aprobaron el Curso de Articulación Disciplinar, del grupo de estudiantes que tienen la formación suficiente para realizar Matemática A o sea que han aprobado el Curso de Articulación Disciplinar. Esta Tabla de resumen, muestra la cantidad de datos discriminados correcta e incorrectamente. En este estudio se tiene que el 96 % se clasifica correctamente.

			Grupo de pertenencia pronosticado		
		discriminante	no	si	Total
Original	Recuento	no	14	5	19
		si	3	176	179
	%	no	73,7	26,3	100,0
		si	1,7	98,3	100,0

Tabla 6: Resultados de la clasificación

Como se puede observar de la tabla, en el grupo correspondiente a alumnos del Curso Tutorial la técnica es menos eficiente en su discriminación (26,3 % de casos mal clasificados). Una de las posibles causas es la diferencia marcada en los tamaños de los grupos. No obstante, se debe aclarar que en el grupo minoritario no se obtuvo respuesta en algunas de las variables, y como se dijo anteriormente la técnica requiere casos con información completa. Se desea repetir este estudio el siguiente año, poniendo especial esfuerzo en obtener datos en el grupo de alumnos que no aprueban el Curso de Articulación Disciplinar.
Si bien son muy pocos los casos mal clasificados de los alumnos que pertenecen al grupo de Matemática A y el modelo los clasifica como pertenecientes al grupo Tutorial (1.7 %), la lectura de esta situación nos es de utilidad para detectar alumnos que cuentan aún con deficiencias en su formación previa en matemática comprometiendo su desempeño en este curso. El uso de estrategias pedagógicas alternativas permitiría subsanar esta situación.

	Coeficientes de las funciones de Clasificación	
Variable	Alumno del Curso Tutorial	Alumno de Matemática A
Estudió Parábola	11.97	14.80
Estudió Trigonometría	3.63	4.43
Nota Curso Articulación	0.58	1.62
Mat. último año	5.49	5.02
Constante	-11.16	-16.51

Tabla 7: Coeficientes para la función discriminante lineal

Fisher propuso una función de clasificación para cada grupo. En el caso de dos grupos la diferencia entre ambas funciones da lugar a un vector de coeficientes proporcional a los coeficientes no estandarizados de la función discriminante canónica. Para clasificar nuevos casos se utilizan estas funciones de clasificación para cada grupo cuyos coeficientes se presentan en la tabla 7. Al utilizarlas, se clasifica un alumno en el grupo para el que la función sea mayor.

Regresión logística

Para cada alumno se desea establecer si tiene una formación lo suficientemente robusta como para realizar el primer curso de las carreras de Ingeniería, Matemática A, o requiere del apoyo especial del Curso Tutorial en Matemática, quedando identificados los alumnos ingresantes en dos grupos. Estos dos grupos pueden verse como una variable dicotómica que considera la pertenencia o no de los alumnos a Matemática A según si aprobaron el Curso de Articulación Disciplinar. Un modelo apropiado cuando la variable dependiente es dicotómica es el modelo de regresión logística. Este tipo de modelo, también llamado modelo de respuesta cualitativa, tiene utilidad para pronosticar qué sucederá cuando existan dos posibilidades, en nuestro caso, superar dicho curso o no. Esto se logra mediante un modelo que calcula probabilidades para cada caso permitiendo asignarlo a un grupo bien definido según si esta probabilidad supera o no un punto de corte previamente fijado (0.65 en nuestro caso).

El análisis de regresión logística [2,6] presenta la ventaja de no requerir que las variables se distribuyan según la ley normal, lo que muchos argumentan como la razón fundamental para que este enfoque resulte notoriamente más robusto que el enfoque discriminante.

Este tipo de modelos permiten estimar o predecir la probabilidad de que un individuo posea una característica en función de determinadas cualidades individuales o variables (x_1, x_2, .., x_k).

En regresión logística el modelo matemático que mejor estima tal probabilidad, debido a que restringe los valores a su rango $0 < P < 1$ y aproxima a una forma ¨S¨ de la curva, es el siguiente: $P(Y=1)=\dfrac{e^{a+bx}}{1+e^{a+bx}}$

o equivalentemente: $P(Y=1)=\dfrac{1}{1+e^{-(a+bx)}}$

Para la estimación de los coeficientes *b* utilizamos el método de máxima verosimilitud.

Para nuestros datos el modelo resultante es el siguiente:

$$P(Z=1)=\frac{1}{1+\exp(-112.84+32.41\text{NotCur}-0.45\text{MatUlt}-0.12\text{EstTrig}+0.69\text{EstPar})}$$

Para ver la de bondad de ajuste del modelo se consideran las pruebas de Hosmer y Lemeshow y la prueba que considera *–2log* de la versosimilitud. Ambos test indican un buen ajuste del modelo aquí presentado.

Observado		Pronosticado[a]		
		discriminante		Porcentaje correcto
		no	si	
discriminante	no	19	0	100,0
	si	0	179	100,0
Porcentaje				100,0

a. El valor de corte es 0,650

Tabla 8: Resultados de la clasificación

Conclusiones

Periódicamente se realizan estadísticas usando las bases de datos de los alumnos ingresantes a los fines de diagnosticar y tomar decisiones en busca de una mejor organización y calidad de enseñanza. Persiguiendo este objetivo es que se ha elegido aplicar técnicas de análisis multivariado que logran identificar un conjunto de variables que caractericen la formación de los alumnos ingresantes en el área matemática. Esto permite obtener una clasificación para ubicarlos en el curso tutorial o en Matemática A. Tanto las técnicas de análisis discriminante como de regresión logística proporcionaron buenas clasificaciones para los casos estudiados (96 y 100 % de los casos correctamente clasificados). Estas técnicas permiten realizar un análisis individual haciendo un diagnóstico sobre la formación actual de cada alumno, evaluando si es apropiada para un buen desempeño en el primer curso de matemática. Cuando la función discriminante aplicada a casos conocidos clasifica incorrectamente un alumno de matemática A y lo coloca en el grupo que debe realizar el curso tutorial, indica que debe dársele especial atención a este alumno debido a que puede evidenciar deficiencia en su formación matemática.

En el futuro, la inclusión de más preguntas tendientes a obtener mayor información, permitirá un diagnóstico más desagregado.

La detección de los grupos con deficiencias permite poner en marcha políticas y estrategias de enseñanza dirigidas a su corrección y la reducción de situaciones de fracaso.

Referencias bibliográficas

[1] Ávila O. y otros, *Rendimiento en Matemática de los alumnos ingresantes en la Facultad de Ingeniería Química, Universidad Nacional del Litoral*. Actas de Congreso, UMA, Salta, Argentina 2005.

[2] Hosmer D. y Lemeshow, *Applied Logistic Regresión*, Estados Unidos, 1989.

[3] Johnson R. y D. Wichern, *Statistical Multivariate Statistical Analysis*, Fourth Edition, Prentice Hall, 1998.

[4] LeBold W. y otros, *The use of discrimnant analysis for optimal placement*, ASEE Annual Conference Proceedings, Purdue University, W. Lafayette, Estados Unidos, 1989.

[5] LeBold W. y otros, *Understanding of Mathematics and Science: Efficient Models for Student Assessments* (Vol.41, Nro 1 IEEE Transactions on Education, Purdue University, W. Lafayette), Estados Unidos, 1998.

[6] Pampel F., *Logistic Regresión*, Estados Unidos, 2000.

(*) Depto de Matemática, Facultad de Ingeniería Química, UNL, Santa Fe.
(**), IMAL, CONICET, Santa Fe.