Entre Ciencia e Ingeniería, año 12 (vol 12), no. 23 – Primer semestre de 2018, página 109 - 116 DOI: http://dx.doi.org/10.31908/19098367.3710
ISSN 1909-8367 (Impreso), ISSN 2539-4169 (En línea)

Tipología influyente en el rendimiento académico de alumnos universitarios[1]

Influential typology in the academic performance of university students

Tipologia influente no desempenho acadêmico de estudantes universitários

S. Ruiz, M. Herrera, M. Romagnano, L. Mallea y M. I. Lund

Recibido: diciembre 10 de 2017 – Recibido: febrero 15 de 2018

***Resumen*—El presente trabajo aborda un informe estadístico centrado en caracterizar el rendimiento académico de alumnos universitarios, a partir de la determinación de variables asociadas, aplicando técnicas estadísticas del Análisis Multivariado. Los análisis efectuados se basan en datos provenientes de una encuesta realizada en el año 2015, a los alumnos de la Facultad de Ciencias Exactas Físicas y Naturales, de la Facultad de Filosofía, Humanidades y Artes de la Universidad Nacional de San Juan, Argentina. Mediante un Análisis Factorial de Correspondencias Múltiples, Análisis de Conglomerados y Análisis de Discriminación Logística, se pudieron identificar tipologías de alumnos y variables influyentes que diferencian a los alumnos según su rendimiento. Los resultados aportan herramientas que permiten realizar un válido diagnóstico para orientar de manera efectiva las intervenciones que realice la institución educativa.**

***Palabras clave*—clasificación, clúster, discriminación, factores, rendimiento, alumnos.**

***Abstract*—The present work addresses a statistical report focused on characterizing the academic performance of university students, from the determination of associated variables, applying statistical techniques of Multivariate Analysis. The analyzes performed are based on data from a survey conducted in 2015, among the students of the Faculty of Exact, Physical and Natural Sciences and the Faculty of Philosophy, Humanities and Arts of the National University of San Juan. Through a Factorial Analysis of Multiple Correspondences, Cluster Analysis and Logistic Discrimination Analysis, it was possible to identify types of students and influential variables that differentiate the students according to their performance. The results contribute to bring tools that allow a valid diagnosis to effectively guide the interventions made by the educational institution.**

***Keywords*—performance, students, classification, discrimination, factors, cluster.**

***Resumo*—O presente trabalho aborda um relatório estatístico centrado na caracterização do desempenho acadêmico de universitários, a partir da determinação de variáveis associadas, aplicando técnicas estatísticas de análise multivariada. As análises realizadas baseiam-se em dados de uma pesquisa realizada em 2015, entre os alunos da Faculdade de Ciências Exatas, Físicas e Naturais e da Faculdade de Filosofia, Humanidades e Artes da Universidade Nacional de San Juan. Por meio de uma Análise Fatorial de Correspondências Múltiplas, Análise de Cluster e Análise de Discriminação Logística, foi possível identificar tipos de alunos e variáveis influentes que diferenciam os alunos de acordo com seu desempenho. Os resultados contribuem pelo fornecimento de ferramentas que permitem um diagnóstico válido para orientar efetivamente as intervenções realizadas pela instituição de ensino.**

***Palavras chave*—desempenho, alunos, classificação, discriminação, fatores, cluster.**

---

[1]Producto derivado del proyecto de investigación "Técnicas de clasificación en el rendimiento académico universitario", aprobado por evaluadores externos y financiado por la Universidad Nacional de San Juan, Argentina.

S. Ruiz, docencia e investigación en el Departamento de Geofísica y Astronomía de la FCEFN - UNSJ, San Juan Argentina,
email: sbruizr@yahoo.com.ar.

M. Herrera, investigación en el Departamento de Informática de la FCEFN - UNSJ, San Juan Argentina, email: mherrera@iinfo.unsj.edu.ar.

M. Romagnano, investigación en el Instituto de Informática de la FCEFN - UNSJ, San Juan Argentina, email: maritaroma@iinfo.unsj.edu.ar.

L. Mallea, docencia en el Departamento de Matemática de la FFHA – UNSJ, San Juan Argentina, email: lamallea@gmail.com.

M. I. Lund, investigación en el Instituto de Informática de la FCEFN - UNSJ, San Juan Argentina, email: mlund@iinfo.unsj.edu.ar.

Como citar este artículo: Ruiz, S., Herrera, M., Romagnano, M., Mallea, L. y Lund, M. I. Tipología influyente en el rendimiento académico de alumnos universitarios, Entre Ciencia e Ingeniería, vol. 12, no. 23, pp. 109-116, enero - junio, 2018.
DOI: http://dx.doi.org/10.31908/19098367.3710

I. Nomenclatura

FCEFN – Facultad de Ciencias Exactas, Físicas y Naturales.

FFHA – Facultad de Filosofía, Humanidades y Artes.

UNSJ – Universidad Nacional de San Juan.

II. Introducción

En muchas investigaciones, independientemente del área de conocimiento, es habitual contar con la necesidad de identificar cuáles son las características que diferencian ciertos grupos de sujetos u objetos respecto de otros, para así contar con predicciones futuras [1].

Este trabajo tiene como propósito esencial mostrar los resultados obtenidos en la determinación de variables que mejor explican la atribución de la diferencia de los grupos de alumnos universitarios de la FCEFN y FFHA de la UNSJ, según su buen o mal rendimiento. Para ello se aplican técnicas del Análisis Multivariado (AM), a datos provenientes de una encuesta titulada: "Encuesta de factores de riesgo y calidad de vida de estudiantes universitarios", realizada a alumnos de dichas facultades.

El Análisis Factorial de Correspondencias Múltiples (AFCM) es una herramienta del AM de gran utilidad en la investigación por encuestas, tanto por su potencial en términos exploratorios como por su adecuación para el tratamiento de variables categóricas. Permite establecer las relaciones (correspondencias) que existen entre las variables (y entre sus modalidades). Puede interpretarse como una manera de representar las variables en un espacio de dimensión menor, mediante la definición de ejes factoriales y utilizando la métrica Chi-cuadrado. También, como un procedimiento objetivo de asignar valores numéricos a variables cualitativas [2].

Por otro lado, tanto el Análisis de Conglomerados como el Análisis Discriminante, lo que algunos autores ubican entre las técnicas estadísticas del AM más potentes para aplicar en investigaciones sociales, son técnicas que permiten clasificar sujetos u objetos a partir de características similares. Las técnicas mencionadas se pueden diferenciar de acuerdo con la forma de extraer conocimiento útil, escondido en esos datos. El Análisis Discriminante cuenta con grupos de datos conocidos, así como observaciones de unidades cuya pertenencia a los grupos, en términos de los grupos conocidos, es desconocida inicialmente y tiene que ser determinada a través del análisis de los datos. Este tipo de problemas de clasificación es comúnmente conocido como reconocimiento de patrones asistido, o aprendizaje con una guía. En terminología estadística, se conoce con el nombre de "Análisis Discriminante" (AD).

De otro lado, existen problemas de clasificación donde los grupos son ellos mismos desconocidos a priori y el principal propósito del análisis es determinar los grupos a partir de los propios datos, de modo que las unidades dentro del mismo grupo sean en algún sentido más similares u homogéneas que aquellas que pertenecen a grupos diferentes. Este tipo de problema de clasificación es referido como reconocimiento de patrón no supervisado o conocimiento sin guía, y, en terminología estadística, se conoce con el título de "Análisis de Conglomerados" (AC).

De otro lado, se puede afirmar que, en general, un indicador directo de la calidad de la enseñanza es el rendimiento académico medido a través del nivel alcanzado por los estudiantes [3]. Así, dicho rendimiento del estudiantado universitario constituye un factor imprescindible en el abordaje del tema de la calidad de la educación superior. Vista la importancia del tema, en este trabajo se aplican y complementan las distintas técnicas mencionadas para analizar y caracterizar lo que se denomina rendimiento académico universitario desde la perspectiva del alumno.

III. Metodología

Teniendo en cuenta que se quiere caracterizar el rendimiento académico hallando tipologías del alumnado, como también determinar variables influyentes y definir una función discriminante que explique el rendimiento de los alumnos universitarios a partir de dichas variables, se organiza el presente trabajo en dos etapas:

1) Caracterización del alumnado según su perfil de rendimiento.

2) Definición y evaluación de modelos que permitan identificar variables influyentes y discriminar a alumnos según su rendimiento.

Para tal efecto, se realiza un estudio exploratorio uni y bidimensional de los datos de la encuesta. Como existe una evidente asociación de las variables tratadas, es necesaria la aplicación de técnicas del AM. La información proveniente de la encuesta incluye preguntas relacionadas con el rendimiento académico del estudiante. La encuesta fue elaborada con la herramienta web de encuestas online, EncuestaFácil.com, y cuenta con varias secciones. Las variables consideradas se pueden agrupar en: variables que caracterizan la Facultad, la universidad y la carrera/s que cursa el estudiante; variables que representan características personales del estudiante y de su familia (edad, sexo, su relación con pares, etc.). Variables asociadas al rendimiento (rendimiento, promedio, etc.); y variables que representan el esfuerzo y motivación del estudiante (Ej. hs. de estudio, asistencia a la universidad, etc.). Se puede consultar la encuesta en:

https://www.encuestafacil.com/RespWeb/Qn.aspx?EID=2197195.

Las preguntas consideradas no son mutuamente excluyentes entre sí, por lo que cada pregunta es una variable en sí misma; las respuestas alternativas a las preguntas de cada sección sí son mutuamente excluyentes y cada una de ellas es una modalidad de las variables cualitativas a las que pertenece. En este trabajo, cada encuestado no interesa en sí mismo sino como representante de cierta categoría o grupo de población. En la primera etapa, mediante el AFCM y un AC, se busca encontrar la tipología de los alumnos según su perfil de rendimiento y su caracterización. Dos alumnos son próximos si poseen buen rendimiento o no, y si tienen similares características según las variables consideradas. Con el AFCM se pretende: hallar la semejanza de los alumnos; estudiar la relación entre las variables; resumir

las características observadas en un pequeño número de variables; y comparar modalidades de diferentes variables.

Se distinguen dos ámbitos: el que concierne al estudio de cómo se agrupan los alumnos teniendo en cuenta su rendimiento académico, y las variables más asociadas. A estas variables se les llama en este artículo: "variables activas" en la conformación de los clúster en el AC. El segundo ámbito es relativo a otras variables, distintas de las activas, a las que denominan los autores: "variables suplementarias", que pueden contribuir con la caracterización de los grupos de alumnos predefinidos, en otros aspectos que pueden ser relevantes.

Una vez que se obtienen los factores que concentran la mayor proporción de inercia, mediante el AFCM, se puede aplicar un AC. Dicho análisis trata, a partir de una tabla de datos (individuos-variables), de situar a los individuos en grupos homogéneos o conglomerados, de manera que los que se puedan considerar similares, sean asignados a un mismo clúster o grupo. Tanto el AFCM como AC se realizan con la ayuda del software estadístico SPAD-N, siguiendo los pasos indicados en la Fig. 1. SPAD (Système Portable pourl'Analyse de Données) permite implementar una estrategia de análisis adecuada al tratamiento exploratorio multivariante de grandes tablas de datos. Su concepción es original y adaptada para un proceso natural de aprendizaje a partir de los datos (data learning) [4]. La existencia de asociación entre variables se puede pensar como una fuente de oportunidades para plantear Modelos de Discriminación. El objetivo es definir un modelo para clasificar mediante la construcción de un Modelo de Regresión Lineal Generalizado (MLG), tal que a partir de una observación "x" de n-variables asociadas al rendimiento académico del alumno, proporcione la probabilidad de que el rendimiento académico sea, por ejemplo, malo (éxito).

Dicho modelo se construye maximizando la probabilidad de la muestra. A través de un Modelo de Regresión Logística (MRL) se puede estimar la probabilidad de un suceso que depende de los valores de ciertas covariables o variables asociadas. En la segunda etapa de este estudio, el suceso (o evento) de interés es *A*: "El Rendimiento Académico del alumno es malo", que puede presentarse o no en cada uno de los alumnos de la población donde se realizó la encuesta. Se considera la variable binaria "*y*" (de tipo Bernoulli) que toma los valores: $y = 1$, si el suceso A se presenta; $y = 0$, si A no se presenta. Sea p la probabilidad de que $y=1$, en un ensayo. Suponiendo que la probabilidad p depende de los valores de ciertas variables, $X_1,\dots,X_k$; si son las observaciones correspondientes a un alumno sobre las variables, entonces la probabilidad de acontecer A dado x es $p(y=1/x)$ que simbolizamos con $p(x)$. La probabilidad de que no suceda A dado x será $p(y = 0/x) = 1\text{-}p(x)$.

Teniendo en cuenta los valores en que varía p(x) y el tipo de variables explicativas en este estudio, resulta conveniente suponer un modelo lineal con la llamada transformación logística de la probabilidad (1), siendo los $1\leq i \leq k$, con los parámetros del MRL.

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + x_1\beta_1 + \cdots + x_k\beta_k \qquad (1)$$

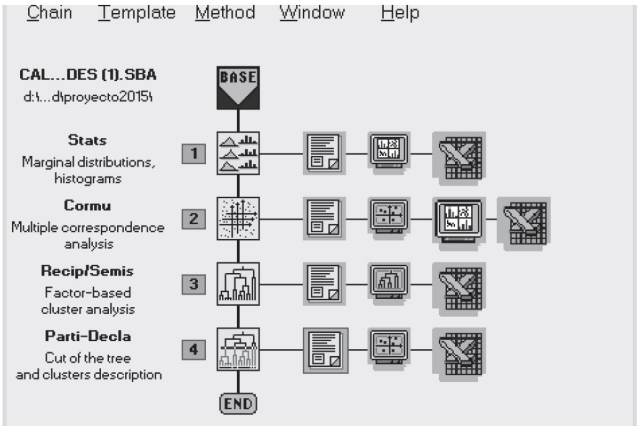


Fig. 1. Pasos en la clasificación jerárquica con SPAD-N.

Según Díaz y Demetrio [5], el objetivo en el proceso de ajuste de un MLG determinado debe ser obtener el mejor trade-off entre el número de variables y sus parámetros, que deben incluirse en la estructura lineal, manteniendo el menor número posible de ellos y la habilidad del modelo para representar a los datos, conservando el ajuste lo más adecuado posible.

Como test de ajuste de un modelo MLG, primeramente se observa el estadístico deviance [6], diferencia entre los máximos de los log-verosimilitud para el modelo saturado y en investigación (modelo con k parámetros), esto es, $D = 2(\hat{L}(n) - \hat{L}(k))$.Otra expresión para este estadístico es $S_k = \frac{D(y,\hat{u})}{\phi}$, conocido como la deviance para el modelo de investigación. Para ver si un modelo ajusta bien, se compara el valorS_k con el punto crítico de la X^2(n-k)gl, para un nivel de significación α. Luego si $S_k > X^2$(n-k)gl;α el modelo de investigación es rechazado. Para analizar la contribución de un término más en el modelo, se emplea la diferencia de deviances, $S_k - S_q$, la cual debe ser comparada con el punto crítico de la X^2(n-k)gl para ese nivel α. También se utiliza la cantidad denotada con AIC, de la forma $D+2k\hat{\phi}$ presentada por Chambers y Hastie, con $\hat{\phi}$ parámetro de escala [7].

Para determinar un MRL se aplica la función genérica "glm" del programa R. Como el modelo de discriminación no es único, es conveniente evaluar y comparar resultados. Para ello se aplica el método Hold-out que tiene en cuenta las tasas de errores en la clasificación (APER), tanto en la muestra para definir el modelo, como en la muestra test. La tasa de error aparente se define como la fracción de observaciones en las muestras de entrenamiento, clasificadas erróneamente por la función de clasificación muestral, determinada por $APER = \frac{n_{1M}+n_{2M}}{n_1+n_2}$, donde n_{1M} es el número de observaciones de la población 1, clasificado erróneamente como observaciones de la población $2yn_{2M}$; el número de observaciones de la población 2 clasificados erróneamente como observaciones de la población 1 [8]. En este estudio, la selección de la muestra para el entrenamiento se determinó de forma aleatoria a partir de una rutina "simple" del programa *R*, y representando aproximadamente el 70% del total de datos de la encuesta.

IV. Desarrollo y Resultados

A partir de la puesta en práctica de una encuesta sobre factores de riesgo y calidad de vida a 74 alumnos universitarios pertenecientes a la FFHA y FCEFN de la UNSJ, donde aproximadamente el 88% afirma tener "buen rendimiento académico", y el 12% "mal rendimiento", se quiere caracterizar a los alumnos en cuanto a su rendimiento, con el objetivo de identificar grupos y diferenciarlos, con base en el análisis de variables asociadas, utilizando las técnicas estadísticas del AC y el Análisis de Discriminación Logística. En la base de datos se pueden distinguir 95 variables; se trabaja con 30 de ellas que se consideran más adecuadas para la caracterización del rendimiento académico del estudiante de ambas facultades; 6 de las variables se seleccionan como "variables activas" para la construcción de los clúster o grupos. Las 24 variables restantes se consideran "variables suplementarias" para caracterizar a los grupos. Las variables seleccionadas son todas categóricas y el número de modalidades o categorías quedan detalladas en la Tabla I.

El criterio de selección de cada variable categórica como activa, fue a través de observar evidencias, bajo un nivel de significación del 5% en la asociación de la variable seleccionada y la variable rendimiento académico. Esto requirió previamente del análisis y la transformación de la base de datos en forma conveniente. Se combinaron categorías de una misma variable cuando las frecuencias observadas resultaron menores a 5, y se aplicaron pruebas Chi-cuadrado de independencia para tablas de contingencias dos por dos. En total se realizaron 29 pruebas utilizando el software SPSS. Estas variables, así seleccionadas, son las que van a intervenir en la definición de los ejes factoriales (Tabla II y Tabla III), en el AFCM.

En las Tablas II y III se muestran algunos resultados del AFCM correspondiente a la primera etapa de trabajo. Del análisis se observa:

• Para el primer eje factorial, las variables activas que más peso ejercen sobre el eje, y concentran mayor cantidad de inercia son: "Tu Rendimiento Académico es" (contribución 23.6); "Tienes Internet en vivienda" (contribución 21.2); "Acceso a PC en vivienda"; y "Relación con compañeros" (las dos últimas con contribución 19). Los valores de los cosenos al cuadrado de cada modalidad, en la medida en que se aproximan al valor 1, indican buena calidad en la representación de la modalidad en el eje, por lo que todas las modalidades quedan bien representadas en el eje. Las modalidades de mayor peso ubicadas en el semieje positivo son: "Bueno" (respecto a la variable "Rendimiento Académico"); "Sí" (tiene Internet en su vivienda); "Sí" (tiene acceso a PC en vivienda), y "Buena" (relación con compañeros). En el primer eje factorial, que reúne el 39.70% de la inercia total de la nube de puntos (ver Tabla II), se oponen los alumnos que: tienen buen rendimiento, PC e Internet en vivienda, y consideran llevar buena relación con sus compañeros; con aquellos que poseen mal rendimiento, no tienen acceso a PC ni a Internet en su vivienda y llevan mala, regular o ninguna relación con sus compañeros.

Tabla I

Caracterización De Variables Activas Y Suplementarias.

SELECCIÓN DE VARIABLES CATEGÓRICAS (6 variables activas y 24 variables suplementarias)		
Variables Activas	Acceso a PC en vivienda	2 categorías
	Tienes Internet en Vivienda	2 categorías
	Valor Alimento	2 categorías
	Cuántas hs a estudiar	2 categorías
	Relación c/compañeros	2 categorías
	Tu Rendimiento académico es	2 categorías
Variables Suplementarias	Año que cursas	6 categorías
	Género	3 categorías
	Edad	3 categorías
	Tienes Hijos	2 categorías
	Cómo costeaste estudios	4 categorías
	Valor Estudio	2 categorías
	Con qué frecuencias vas a Universidad	4 categorías
	Promedio con Aplazo/s	5 categorías
	Importancia de Aplazo/s	6 categorías
	Nivel exigencia carrera	3 categorías
	Cuántas horas dedicas a clase Teórica	4 categorías
	Cuántas horas dedicas a la clase Práctica	4 categorías
	Cuántas horas dedicas a la consulta	4 categorías
	Material Estudio Económicamente accesible	3 categorías
	Material Estudio Actualizado	3 categorías
	Material Estudio de Calidad	3 categorías
	Material de fácil comprensión	3 categorías
	Material de Biblioteca	5 categorías
	Elección de la carrera	2 categorías
	Relación con el docente	4 categorías
	Estrés por estudiar	4 categorías
	La carrera mejora tu futuro	3 categorías
	Horas al día sentado	4 categorías
	Percepción de Salud	7 categorías

- En el factor 2 la variable relevante es "Cuántas horas le dedica a estudiar" (contribución 63.4). Teniendo en cuenta los valores de los cosenos al cuadrado, las dos modalidades de dicha variable quedan bien representadas. El segundo eje, que reúne el 18.88% de la inercia total (ver Tabla II), contrapone los alumnos que le dedican más de 4 horas de estudio diario a la carrera (representado en el semieje positivo), respecto a aquellos que le dedican menos (representado en el semieje negativo).
- El tercer eje factorial, que reúne el 14.83% de la inercia total, la variable más relevante es "Valor alimento", con una contribución 61.2.

Tabla II
Caracterización De Los Ejes Factoriales.

N° de eje	Eigenvalor	Porcentaje	Porcentaje acumulado
1	0.3970	39.70	39.70
2	0.1888	18.88	58.59
3	0.1483	14.83	73.42
4	0.1097	10.97	84.39
5	0.0916	9.16	93.54
6	0.0646	6.46	100.00

Tabla III
Contribuciones de las Variables Activas Según los 5 Primeros Ejes Factoriales.

Variables Activas	Peso Relativo	Distancia al origen 1	Eje		
			2	3	
Acceso a PC en vivienda					
No	1.802	8.25	16.96	20.65	9.03
Sí	14.865	0.12	2.06	2.50	1.09
Tienes internet en vivienda					
No	4.505	2.70	15.49	5.36	6.31
Sí	12.162	0.37	5.74	1.99	2.34
Valor alimento					
Suficiente	14.865	0.12	1.39	0.01	6.61
Insuficiente	1.802	8.25	11.45	0.04	54.57
Cuántas horas a estudiar					
Más de 4	12.838	0.29	1.01	14.57	0.01
Menos de 4	3.829	3.35	3.37	48.84	0.03
Relación c/compañeros					
Buena	14.189	0.17	2.82	0.43	2.10
Reg-Mala-Inexistente	2.477	5.72	16.15	2.46	12.01

Teniendo en cuenta las proyecciones de calidad en el primer plano factorial de las modalidades de las variables suplementarias, las cuales se pueden identificar a través de los valores-test mayores a 1.96 (en valor absoluto), que conducen a rechazar la hipótesis aleatoriedad con un nivel de significación de 0.05, se observa que: las variables suplementarias relacionadas con el primer eje factorial son "Año que cursas", "Edad", "Cómo costeaste estudios", "Promedio con aplazos", "Valor Estudio", "Importancia de los aplazos", "Cuántas horas le dedica a la práctica", "Material de Biblioteca" y "Relación con el docente".

Los alumnos cuyo rendimiento académico es bueno, tienen PC e Internet en vivienda, suficiente dinero para alimento y estudio, dedican más de 4 horas diarias a estudiar, tienen buena relación con compañeros y docentes, cursan tercer año, su edad oscila entre los 18 y los 24 años, costean los estudios solo con aporte, y el promedio con aplazos es mayor a 8. Se oponen a aquellos que poseen mal rendimiento, no tienen PC ni Internet en vivienda, es insuficiente el dinero para alimento y estudio, le dedican menos de 4 horas a la práctica y a estudiar, expresan que su relación con compañeros es regular, mala o inexistente, cursan cuarto año, la edad oscila entre 24 a 29 años, costean el estudio solo con trabajo, sienten mucho fracaso cuando lo aplazan, consideran que el material de Biblioteca es de difícil acceso y su relación con docentes es regular.

Mediante el AC se clasifican los alumnos de acuerdo con su similitud, por clasificación jerárquica, utilizando el método del vecino más próximo, considerando los primeros tres ejes factoriales que reúnen una inercia total acumulada del 73,42% aproximadamente. A partir de ella se obtiene la representación gráfica del proceso del agrupamiento mediante un Dendograma, en la que se pudieron distinguir dos grupos.

La primera clase o grupo la componen 12 alumnos, que representan el 6,22% del total de los alumnos encuestados. Se destacan por: no poseer internet en vivienda (55% del total de alumnos de esta modalidad forman esta clase, y el 91.67% de los alumnos de la clase no tienen internet en su casa); no tienen acceso a PC en vivienda (87.50% del total de alumnos de esta modalidad forman esta clase, y el 58.33% de los alumnos de la clase no tienen acceso a PC en vivienda); su rendimiento académico es malo (77.78% del total de los alumnos que presentan esta modalidad; representa el 58.33% de los alumnos de la clase no tienen buen rendimiento); su relación con compañeros es regular, mala o inexistente (63.64% del total de los alumnos que presentan esta modalidad; representa el 58.33% de los alumnos de la clase tienen esta modalidad); el dinero para alimento y estudio considera que es insuficiente (75%, 28.57% del total de alumnos que presentan cada modalidad; 50%, 83.33% de los alumnos de la clase tienen la modalidad, respectivamente); cursan cuarto año (71.43% del total de los alumnos que presentan esta modalidad; representa el 41.67% de los alumnos de la clase tienen esta modalidad); Cuentan entre 24 a 29 años de edad (36% del total de los alumnos que presentan esta modalidad; representa el 75% de los alumnos de la clase tienen esta modalidad); sienten mucho fracaso con los aplazos (31.03% del total de los alumnos que presentan esta modalidad; representa el 75% de los alumnos de la clase tienen esta modalidad); y considera que su relación con los docentes es regular (36.84% del total de los alumnos que presentan esta modalidad; representa el 58.33% de los alumnos de la clase tienen esta modalidad).

La segunda clase reúne el 83.78% del total de los alumnos encuestados, que se caracterizan por: poseer Internet en vivienda (98.15% del total de alumnos de esta modalidad; 85.48% de los alumnos de la clase); tener acceso a PC en vivienda (92.42% del total de alumnos de esta modalidad; 98.39% de los alumnos de la clase); tener buen rendimiento (92.31% del total de alumnos de esta modalidad; 96.77% de los alumnos de la clase); su relación con compañeros es buena (92.06% del total de alumnos de esta modalidad; 93.55% de los alumnos de la clase); consideran que tienen suficiente dinero para alimento y estudio (90.91%, 94.87% del total de alumnos que presentan cada modalidad; 96.77%, 59.68% de los alumnos de la clase tienen la modalidad, respectivamente); costean los estudios solo con aporte (100% del total de alumnos de esta modalidad; 43.55% de los alumnos de la clase); su edad está entre los 18 y los 24 años

(95.12% del total de alumnos de esta modalidad; 62.90% de los alumnos de la clase); y su promedio con aplazo es mayor a 8 (100% del total de alumnos de esta modalidad; 35.48% de los alumnos de la clase).

En la segunda etapa de trabajo, para definir un modelo de discriminación mediante MRL, se dispone de la información proveniente de la encuesta, para las variables categóricas cuyas modalidades se definen en la Tabla IV.

Tabla IV
Variables Consideradas para Definir El MRL

Lista de variables	
V01 Rendimiento académico **0 Bueno** **1 Malo** **V05 Tiene PC en vivienda** **0 No PC vivienda** **1 Sí PC vivienda** **V06 Tiene internet vivienda** **0 No internet vivienda** **1 Sí internet vivienda**	V07 Dinero para Alimento 0 Suficiente alimento 1 Insuficiente alimento V15 Horas que dedica a estudio 0 Más de 4 hs 1 Menos de 4 hs V20 Relación con Compañeros 0 Buena relación 1 Regular-Mala-inexistente

Se confecciona un programa en *R* cuyos pasos se pueden resumir en los siguientes:

1. Lectura de la base de datos.
2. Categorización de las variables explicativas (V05, V06, V07, V15 y V20).
3. Categorización de la variable respuesta (V01).
4. Selección de una muestra de entrenamiento y una muestra test.
5. A partir de la muestra de entrenamiento se realiza la estimación de los parámetros, y las pruebas de ajuste para el modelo con todas las variables y el modelo de investigación.
6. Confección de tablas de datos mal clasificados, para la muestra de entrenamiento y test.
7. Cálculo de tasas aparentes asociadas a la muestra de entrenamiento y test.
8. Cálculo de odds radio al modelo reducido para interpretar los coeficientes del modelo.

En la parte izquierda de la Tabla V, se observa el resultado de realizar el ajuste por máxima verosimilitud con el software R, considerando todas las variables de la Tabla IV. Los contrastes sobre los parámetros se realizan a través de tests de Wald. La función "glm" devuelve automáticamente el estadístico y la significación del test de Wald para cada parámetro del modelo.

Como la deviance residual S_k=23.966 es menor que el valor crítico $X^2_{(n-k)gl;\alpha} = X^2_{47gl;a=0.05}$=64.00111, se puede asumir que el modelo ajusta bien, con un nivel de significación α=0.05. Luego se aplica el método paso a paso hacia atrás, para analizar si se puede hallar un modelo más simple que el antes fijado. Para ello se utiliza la rutina step de R. Inicialmente se tiene un AIC=33.97 para el modelo con todas las variables, y la función de R considera la eliminación de cada una de las variables para finalmente sacar del modelo la variable V06: "Tiene internet en su casa" por ser la que produce un AIC más bajo (32.16).

Tabla V
Resultados del ajuste del modelo con todas las variables (izquierda) y del modelo reducido (derecha), utilizando la rutina glm de R.

Modelo completo	**Modelo reducido**
Glm (formula=V01~V05+V06+V07+V20, family = "binomial", data = d, subset = train)	glm (formula=V01~V05+V07+V20, family="binomial", data=d, subset=train)
Coeficientes: **Estim. ErrorStd z-valor p-valor** **(Interc) -0.6908 1.2829 -0.538 0.5903** **V05 -2.4386 1.4714 -1.657 0.0975** **V06 -0.6187 1.3984 -0.442 0.6582** **V07 1.8916 1.3979 1.353 0.1760** **V20 2.3674 1.1389 2.079 0.0377*** **Cod-Sig: 0'***' 0.001'**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**	Coeficientes: Estim. ErrorStd z-valor p-valor (Interc) -0.9256 1.1682 -0.792 0.4282 V05 -2.7748 1.2692 -2.186 0.0288* V07 2.2377 1.1644 1.922 0.0546. V20 2.4573 1.1104 2.213 0.0269* Cod-Sig.: 0'***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Deviance Nula: 41.087 con 51 gl.	Deviance Nula: 41.087 con 51 gl.
Deviance Residual: 23.966 con 47 gl.	Deviance Residual: 24.156 con 48 gl.
AIC: 33.966	AIC: 32.156

En el siguiente paso se considera la eliminación de alguna de las tres restantes variables, pero el algoritmo en R decide quedarse con ellas ya que su eliminación supone un aumento, en el mejor de los casos, del AIC último hallado. Como los valores de AIC no difieren mucho, entre el modelo con todas las variables y el modelo reducido, se analiza además la inclusión o no de la variable V06 realizando una prueba de significación mediante una prueba de razón de verosimilitudes. A partir de este modelo bajo un nivel de significación del 5%, como G=Deviance (modelo reducido) – Deviance (modelo con todas las variables)=25.881-25.867=0.014, resulta menor que el valor crítico X^2_{1gl} = 3.841459; por lo tanto, se puede considerar que la inclusión de la variable V06 no contribuye al modelo.

De acuerdo con los resultados más a la derecha en la Tabla V, se pueden observar los parámetros estimados y su significación en el modelo reducido. El MRL reducido se observa en (2).

$$\ln\left(\frac{p}{1-p}\right) = -0.9256 - 2.7748d_{V05} + 2.2377d_{V07} + 2.4573d_{V20} \quad (2)$$

Donde p representa la probabilidad de que la variable V01: "Rendimiento Académico" tome la categoría 1 ("Mal Rendimiento") d_{V05}, d_{V07} y d_{V20}, variables dummy que toman el valor 1 cuando la modalidad que se observa corresponde

a la categoría 1; en caso contrario, las variables toman el valor 0.

Sea $\widehat{\beta_0} + \mathrm{X_i}.\hat{\beta}$ el segundo miembro del modelo reducido planteado, donde $X_i = (\mathrm{d_{V05}}, \mathrm{d_{V07}}, \mathrm{d_{V20}})$, $\widehat{\beta_0} = -0.9256$ y $\hat{\beta} = (-2.7748, 2.2377, 2.4573)^T$ vector de parámetros estimados. Si el *"Alumno i"* tiene x_i como vector de observaciones asociadas a las variables del modelo, $\widehat{p_{(x_i)}} = \frac{1}{1+e^{\widehat{\beta_0}+X_i.\hat{\beta}}} \Rightarrow \widehat{q_{(x_i)}} = 1 - \widehat{p_{(x_i)}} = \frac{e^{\widehat{\beta_0}+X_i.\hat{\beta}}}{1+e^{\widehat{\beta_0}+X_i.\hat{\beta}}}$. se definen como reglas para la clasificación: "Si $\frac{e^{\widehat{\beta_0}+X_i.\hat{\beta}}}{1+e^{\widehat{\beta_0}+X_i.\hat{\beta}}} > 0.5$, entonces el Alumno *i* pertenece a la clase de los de *"Buen Rendimiento"*, caso contrario a los de *"Mal Rendimiento"*.

La Tabla VI representa la tabla de mal clasificados para la muestra de entrenamiento y muestra test, al aplicar el modelo reducido estimado con R. A partir de la tabla se pueden calcular los errores aparentes correspondientes, resultando: 0.07692308 y 0.0909091 respectivamente. Por lo que el porcentaje de mal clasificados para la muestra de entrenamiento es de 7,69%, y para la muestra test de 9,09%. Ambos errores resultan similares, conservando un equilibrio en la proporción de errores, donde supera levemente el error aparente de la muestra test sobre el de la muestra de entrenamiento.

Los coeficientes estimados del modelo reducido se pueden interpretar mediante cocientes de chances. Para ello, se calculan los exponenciales de los coeficientes estimados y sus valores recíprocos en caso de que los exponenciales resulten valores menores a uno. De lo que resulta: si las modalidades "posee suficiente dinero para alimento" y "tiene buena relación con sus compañeros" se mantienen fijas, la chance de que un alumno tenga mal rendimiento, respecto a que tenga buen rendimiento, es $1/e^{-2.7748} \cong 16$ veces mayor, si el alumno no tiene acceso a PC en casa, respecto al que sí tiene acceso.

Ahora, si se mantienen fijas "no tiene acceso PC en su casa", y "tiene buena relación con sus compañeros", la chance de que un alumno no tenga buen rendimiento, respecto a que sí lo tenga, es $e^{2.2377} \cong 9.371751$ veces mayor si "no tiene suficiente dinero para alimento" respecto al que tiene suficiente alimento. Por último, si se mantienen fijas "no tiene acceso a PC en su casa" y "tiene suficiente dinero para alimento", la chance de que el alumno no tenga buen rendimiento, respecto al que sí lo tenga, es $e^{2.4573} \cong 11.67325$ veces mayor si el alumno tiene una relación regular, mala o no tener relación con sus compañeros, respecto al que tiene una relación buena.

V. Conclusiones

Mediante un análisis discriminante se pudo establecer el poder explicativo y discriminatorio de características que diferencian a los alumnos según su rendimiento, a partir de datos extraídos de una encuesta.

A través de la identificación de variables influyentes, mediante la aplicación de técnicas del AM, se pudieron vislumbrar aspectos sociales, culturales y económicos fuertemente asociados al rendimiento del estudiante universitario, y con ello a la calidad educativa, tales como: el disponer de PC en vivienda, el tipo de relación con sus compañeros, tener suficiente dinero para alimento.

Esta información relevante de los alumnos puede contribuir en la formulación de políticas de mejoramiento o direccionamiento institucional. Los resultados aportan herramientas que permitan realizar un diagnóstico válido para orientar de manera efectiva las intervenciones que realice la institución, en este sentido, para posteriormente diseñar una medida de seguimiento año a año mediante la aplicación sistemática del instrumento, con el objeto de evaluar el impacto de las acciones realizadas.

Referencias

[1] Torrado-Fonseca, M. M., Berlanga-Silvente, V. "Revista d'innovació i recerca en educació.," *Rev. d'Innovació i Recer. en Educ.*, vol. 6, no. 2, pp. 150–166, 2013.

[2] Peña, D. *Análisis de datos multivariantes*. McGraw-Hill/ Interamericana, 2002, 515p.

[3] Escudero Escorza, T. "La evaluación y mejora de la enseñanza en la Universidad: otra perspectiva," En *Revista de Investigación Educativa*, vol. 18, no. 2, 2000, pp. 405–416.

[4] Bertaut, M.V., Valls, J.M. "Manual de introducción a los métodos factoriales y clasificación con SPAD." Barcelona, Servei d'Estadística Universitat Autònoma de Barcelona, p. 68.

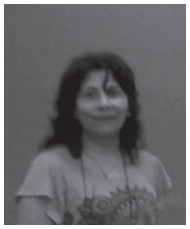
[5] Díaz, M.P., Demétrio, C.G.B. "Introducción a los modelos lineales generalizados: su aplicación en las Ciencias Biológicas". SCREEN Editorial, 1998.

[6] Nelder, J.A., Wedderburn, R.W.M. "Generalized Linear Models," *Source J. R. Stat. Soc. Ser. A J. R. Stat. Soc. A*, vol. 135, no. 3, pp. 370–384, 1972.

[7] Chambers, J., Hastie, T., Pregibon, D. "Statistical Models in S," in *Compstat*, Heidelberg: Physica-Verlag HD, 1990, pp. 317–321.

[8] Richard, R.A., Johnson, A., Wichern, D.W. *Applied multivariate statistical analysis*. Pearson Prentice Hall, 2007.

Susana B. Ruiz. Profesora de Enseñanza Media y Superior en Matemática (1988), Especialista en Docencia Universitaria (1999) de la UNSJ. Magister en Estadística Aplicada (2011) de la Universidad Nacional de Córdoba (Argentina). Desde el año 2016 es profesora Titular Exclusiva en asignaturas del Área de Matemática, del Departamento de Geofísica y Astronomía de la Facultad de CEFyN de la UNSJ. Por extensión cumple tareas docentes en la Cátedra de Teorías de Autómatas y Computabilidad del Dpto. de Informática, y tareas de investigación en el Instituto de Informática de la FCEFyN de la UNSJ. Ponentes en congresos nacionales e internacionales y artículos en revistas en las temáticas: experiencias didácticas a nivel superior, estadística aplicada, estimación de transformaciones con datos contaminados y aplicaciones de Lógica Difusa. Actualmente investigadora en las temáticas: Educación. Estadística aplicada a datos de encuestas. Inteligencia computacional utilizando Lógica Difusa. Minería de Datos.



Myriam Beatriz Herrera. Profesora de Matemática, Magíster (2000) y Doctor en Cs Matemáticas, en la Universidad Nacional de San Luis. (2007). Desde el año 1988 es docente Profesora Adjunta Exclusiva en las asignaturas Probabilidades y Estadísticas y Matemática Básica y desde 2014 Profesora Titular Exclusiva, en el Departamento de Informática de la UNSJ. Ha publicado trabajos en revistas especializadas. Ponente en congresos nacionales e internacionales. Orienta cursos referidos a educación en estadística, estadística teórica y sus aplicaciones.

Autora de libros en el área estadística. Investigaciones en: Reconocimiento de patrones. Procesamiento Estadístico de Imágenes y Estadística Espacial. Minería de datos.

María R. Romagnano. Docente/Investigador de la Universidad Nacional de San Juan, Argentina. Licenciada en Sistemas de Información en el año 2002, en la Universidad Nacional de San Juan. Magister en Informática en el año 2010 de la Universidad de la Matanza. Actualmente es doctoranda del Doctorado en Ingeniería de la Universidad Nacional de Cuyo. Desde el año 2017 es coordinadora titular del Gabinete de Sistemas de Información del Instituto de Informática. Sus intereses en investigación abarcan las áreas de Inteligencia Artificial e Ingeniería de Software.

Lilian Adriana Mallea. Profesora Titular Exclusiva desde el año 2000, en la cátedra Probabilidades y Estadística del Departamento de Matemática-FFHA. Magister (2000) y Doctor en Cs Matemáticas (2007), de la Universidad Nacional de San Luis. Ha publicado trabajos en revistas especializadas, nacionales e internacionales, y ha dictado numerosos cursos en congresos, abordando temáticas referidas a educación en estadística, estadística teórica y sus aplicaciones. Autora de libros en el área estadística. Investigaciones en: Procesos Estocásticos. Procesamiento Estadístico de Imágenes, Estadística Espacial y Minería de datos.

María Inés Lund. Licenciada en Informática (1991), Especialista en Sistemas de Información para Intranets (2002), Magister en Informática (2012). Becaria de investigación de CONICET y Consultora profesional externa. En el año 1991 ingresó como profesora en la UNSJ y desde 2005 es Profesora Adjunto Exclusivo en el Instituto de Informática con extensión a la docencia universitaria. Es Directora de la carrera Licenciatura en Sistemas de Información del Dpto de Informática desde el año 2012 y Coordinadora del Gabinete de Ingeniería de Software del Inst de Informática desde el año 2002. Dirige proyectos de investigación, becarios de investigación, alumnos tesistas de grado y posgrado. Ha realizado presentaciones en congresos nacionales e internacionales, y ha publicado artículos en revistas especializadas.