Detección temprana del rendimiento académico de estudiantes universitarios de primer ciclo mediante el análisis discriminante

Lutzgardo Saavedra*, Julio C. Ramos*, Máximo C. Mitacc*, Víctor R. del Águila*
Universidad de Lima. Perú

Recibido: 29 de marzo del 2017 / Aprobado: 22 de mayo del 2017

Resumen: En este estudio se ha procurado identificar a los ingresantes que aprobarían a lo más dos de los cinco cursos en los que se matricularon para el semestre 2016-2 del Programa de Estudios Generales de la Universidad de Lima. Dicha identificación se basó en modelos de predicción, construidos con datos del semestre 2016-1 mediante el uso de análisis discriminante. La población de ingresantes se dividió en tres dominios de estudio y se construyeron modelos independientes de predicción para el rendimiento académico utilizando las funciones de clasificación de Fisher, evaluadas mediante los indicadores de rendimiento y la curva Receiver Operating Characteristic (ROC).

Palabras clave: rendimiento académico / análisis discriminante / predicción / métodos estadísticos

Early detection of the academic performance of university students from first year through discriminant analysis

Abstract: This work aims to identify students who would only pass at most two out of five enrolled courses from semester 2016-2 of the General Studies Program of Universidad de Lima. The study is based on predictive models constructed with data collected on semester 2016-1 through discriminant analysis. The student's population was divided in three domains of study. Then, independent predictive models for academic performance were constructed using Fisher's classification functions which were evaluated by performance indicators and the Receiver Operating Characteristic curve (ROC).

Keywords: academic performance / discriminant analysis / prediction / statistical methods

* Correos electrónicos: lsaavsan@ulima.edu.pe, jramos@ulima.edu.pe, mmitacc@ulima.edu.pe, vaguila@ulima.edu.pe

1. INTRODUCCIÓN

Actualmente, en la Universidad de Lima el conocimiento sobre el rendimiento académico de los ingresantes es insuficiente por dos motivos principales: en primer lugar, aunque se cuenta con algunos indicadores del rendimiento escolar, muchos no están disponibles digitalmente y otros se encuentran dispersos en aplicaciones informáticas del entorno universitario. En segundo lugar, no se ha implementado un sistema de información para efectuar un estudio estadístico mediante el uso de técnicas predictivas de análisis multivariado, redes neuronales o árboles de decisión. Estas técnicas sistematizarían los datos suficientes para formar una visión global del rendimiento académico de los estudiantes de la universidad con el objetivo de dar soporte científico al proceso enseñanza-aprendizaje, acorde con la política de gestión de calidad.

Con el propósito de solucionar este problema, la Universidad de Lima –por intermedio de la Dirección de Estudios Generales– creó una comisión académica, denominada Análisis de Indicadores. Esta comisión decidió diseñar modelos estadísticos de predicción para automatizar la detección temprana de ingresantes con bajo rendimiento académico. De esta manera, sería posible la implantación oportuna del programa de consejería del Programa de Estudios Generales.

El concepto de rendimiento que se presenta en este trabajo de investigación es el de rendimiento académico como resultado. Según Bloom (1976), consiste en "las diferentes formas que se emplean en cada etapa o nivel de aprendizaje escolar que se toman como base para decidir si pueden pasar a la etapa siguiente".

Cuando nos referimos al rendimiento académico, debemos considerar la definición de Pizarro (como se citó en Reyes, 2003): "una medida de las capacidades respondientes o indicativas que manifiestan, en forma estimativa, lo que una persona ha aprendido como consecuencia de un proceso de instrucción o formación. El mismo autor ahora desde una perspectiva del alumno, define el rendimiento como la capacidad respondiente de este frente a estímulos educativos, susceptible de ser interpretado según objetivos o propósitos educativos preestablecidos".

Como resultado de la naturaleza multidimensional del concepto en cuestión, siempre ha sido difícil establecer el tipo o la medida de rendimiento. En este estudio se adopta un enfoque que combina los aportes de varios autores, a fin de contar con una definición al mismo tiempo amplia y detallada.

Di-Gresia, Porto y Ripani (2002) señalan que la medición del rendimiento de los estudiantes, en cualquier nivel de enseñanza, ha preocupado a investigadores de varias disciplinas. Se considera que los resultados de las pruebas de evaluación o los promedios de notas obtenidas en las materias rendidas o aprobadas pueden ser un primer indicador. Otros factores importantes que influyen en el rendimiento académico son el número de materias aprobadas por año, la cantidad de créditos obtenidos en ese periodo de tiempo, el porcentaje de asistencia, el nivel socioeconómico de los padres de familia, los indicadores socioindividuales del alumno, entre otros.

Además de incluir la medición del rendimiento académico y los factores que influyen en este, el presente estudio contempla algunos antecedentes sobre el tema:

- "Análisis del rendimiento académico mediante un modelo logit" (Ibarra y Michalus, 2010). En este trabajo se analiza el rendimiento académico de los estudiantes de la Facultad de Ingeniería de la Universidad Nacional de Misiones; la población objetivo estuvo conformada por los alumnos de las cohortes de 1999 a 2003 (589 estudiantes). En la investigación se define al rendimiento académico como el promedio de materias aprobadas anualmente, y mediante la técnica estadística multivariada de regresión logística se determina la incidencia de factores de índole personal, socioeconómica y académica. Los resultados obtenidos permiten concluir que las variables significativas del rendimiento académico son tres: el promedio de calificaciones del nivel medio, el tipo de institución donde se cursaron estos estudios y el número de asignaturas aprobadas en el primer año de carrera. Este último factor resultó el más relevante; además, destacó la importancia de la primera etapa de la carrera en los posteriores resultados académicos del estudiante.
- "Aplicación del análisis discriminante para explorar la relación entre el examen de Instituto Colombiano para el Fomento de la Educación Superior (ICFES) y el rendimiento en álgebra lineal de los estudiantes de Ingeniería de la Universidad Tecnológica de Pereira (UTP) en el periodo 2001-2003" (Carvajal, Trejos, y Soto Mejía, 2004). En el artículo se explora la relación entre los puntajes de ICFES en cada una de las áreas del conocimiento y el rendimiento en Álgebra Lineal de los estudiantes de Ingeniería de la UTP, a través de la técnica estadística multivariada, conocida como análisis discriminante. Los datos históricos que sustentan el análisis fueron proporcionados por

la Oficina de Planeación de la UTP. Para el análisis de los datos se utilizó el *software* estadístico SPSS (Statistical Package for Social Sciences, versión 11.5).

El análisis estadístico multivariado desarrollado en este artículo permite concluir que los puntajes obtenidos en las pruebas del ICFES, como único criterio de ingreso a la UTP, dejan serios cuestionamientos sobre su capacidad de prever el desempeño futuro del estudiante en los programas de Ingeniería.

A partir de los antecedentes y con miras a optimizar el conocimiento del rendimiento académico en la Universidad de Lima, en este trabajo se propone detectar a los estudiantes que, a lo sumo, aprobarían dos de los cinco cursos en los que se matricularon en el semestre 2016-2 del Programa de Estudios Generales, sobre la base de los modelos de predicción construidos con los datos del semestre 2016-1 mediante el uso del análisis discriminante. Con este fin, el artículo se organiza en cuatro secciones: introducción, metodología, resultados y conclusiones.

2. METODOLOGÍA

Como se mencionó en el apartado anterior, el objetivo del estudio ha sido identificar a los alumnos que, a lo más, aprobarían dos de los cinco cursos en que se matricularon en el semestre 2016-2 del Programa de Estudios Generales de la Universidad de Lima. El momento idóneo para la identificación fue el inmediatamente posterior al examen parcial, pues entonces aún era posible alguna intervención que mejorara el rendimiento académico de este grupo de ingresantes. A esas alturas del semestre 2016-2, el análisis discriminante permitió construir un modelo con las calificaciones de los estudiantes del semestre 2016-1, capaz de establecer las predicciones utilizando solamente las del semestre 2016-2 obtenidas hasta la mitad del periodo académico en mención.

2.1 Datos

Los datos fueron proporcionados por la Dirección Universitaria de Registros Académicos y Servicio al Programa de Estudios Generales de la Universidad de Lima a solicitud de la Comisión de Indicadores. El universo del estudio se limitó a los datos de los que ingresaron por las modalidades vigentes para los semestres 2016-1 y 2016-2 y excluyó a quienes ingresaron por las modalidades de traslado externo y bachillerato con convalidación.

Para el semestre 2016-1, el conjunto de 2305 registros contenía los siguientes campos: código de ingresante, código de carrera, total de cursos desaprobados en el examen parcial, promedio ponderado de las notas del examen parcial, nota del examen parcial, porcentaje de inasistencias, nota obtenida de las evaluaciones de la tarea académica hasta el examen parcial y nota del examen final, de cada uno de los cinco cursos en los que se matriculó el ingresante. Para el semestre 2016-2, el número de registros fue de 958 con la misma estructura de campos, pero con la columna "Nota del examen final" vacía.

Luego, se inspeccionó el conjunto de datos buscando registros incompletos que podrían distorsionar el análisis. En este proceso, para el semestre 2016-1 se encontraron y eliminaron registros de ingresantes que llevaron menos de cinco cursos. Así, el número de registros se redujo a 2301. En cambio, para el semestre 2016-2 no se detectaron registros incompletos.

Antes de la elaboración del modelo de predicción, los datos de los semestres 2016-1 y 2016-2 fueron agrupados en tres dominios de estudio, los que se muestran en las tablas 1 y 4.

Tabla 1
Distribución de ingresantes en el semestre 2016-1 por dominio de estudio

Dominio de estudio	Número de ingresantes
1. Carreras de Economía, Administración, Marketing, Negocios Internacionales, Contabilidad, Ingeniería Industrial, Ingeniería de Sistemas	1583
2. Carreras de Psicología, Derecho, Comunicación	585
3. Carrera de Arquitectura	133
Total	**2301**

Fuente: Universidad de Lima, Dirección Universitaria de Servicios Académicos y Registro (2016)
Elaboración propia

Las variables propuestas para la elaboración del modelo de predicción fueron las siguientes:

- Variable dependiente
  Definida como número de cursos aprobados por el estudiante en el semestre 2016-1. Esta variable establece las siguientes categorías o grupos:

  - Grupo 0. Estudiantes con a lo más dos cursos aprobados
  - Grupo 1. Estudiantes con tres o más cursos aprobados

- Variables independientes
  Son las variables predictoras que influyen en la variable dependiente. Pueden ser cuantitativas o cualitativas en escala ordinal. Las variables consideradas se muestran en la tabla 2.

Tabla 2
Descripción de las variables, según dominios de estudio

<table>
<tr><th>Dominio 1</th><th>Dominio 2</th><th>Dominio 3</th></tr>
<tr><td colspan="3">X_1: nivel de modalidad de ingreso
X_2: número de cursos desaprobados en el examen parcial
X_3: promedio ponderado acumulado (PPA) del examen parcial
X_4: nota del examen parcial de Lenguaje y Comunicación 1 (LC1)
X_5: porcentaje de inasistencias a LC1
X_6: nota de la práctica calificada de LC1
X_7: nota de la práctica de aula de LC1
X_8: nota del trabajo 1 de LC1
X_9: nota del examen parcial de Desarrollo y Personal Social (DPS)
X_{10}: porcentaje de inasistencias a DPS
X_{11}: nota del trabajo 1 de DPS</td></tr>
<tr><td colspan="2">X_{12}: nota del examen parcial de Metodología de la Investigación (MI)
X_{13}: porcentaje de inasistencias a MI
X_{14}: nota del trabajo 1 de MI
X_{15}: nota del examen parcial de Globalización y Realidad Nacional (GRN)
X_{16}: porcentaje de inasistencias a GRN
X_{17}: nota del trabajo 1 de GRN</td><td rowspan="2">X_{12}: nota del examen parcial de Matemática para Arquitectura (MA)
X_{13}: porcentaje de inasistencias a MA
X_{14}: nota de la práctica calificada 1 de MA
X_{15}: nota de la práctica de aula 1 de MA
X_{16}: nota del examen parcial de Dibujo 1 (D1)
X_{17}: porcentaje de inasistencias a D1
X_{18}: nota del trabajo 1 de D1
X_{19}: nota del examen parcial de Proyecto de Arquitectura 1 (PA1)
X_{20}: porcentaje de inasistencias a PA1
X_{21}: nota del trabajo 1 de PA1</td></tr>
<tr><td>X_{18}: nota del examen parcial de Matemática Básica (MB)
X_{19}: porcentaje de inasistencias a MB
X_{20}: nota de la práctica calificada 1 de MB
X_{21}: nota de la práctica de aula 1 de MB</td><td>X_{18}: nota del examen parcial de Fundamentos de Matemática (FM)
X_{19}: porcentaje de inasistencias a FM
X_{20}: nota de la práctica calificada 1 de FM
X_{21}: nota de la práctica de aula 1 de FM</td></tr>
</table>

Nota: Aunque tienen la misma notación (X_i), las variables son usadas en las ecuaciones de dominios diferentes.
Elaboración propia

Los niveles de modalidad de ingreso X_1 son nueve y se presentan en la tabla 3.

Tabla 3
Niveles de modalidad de ingreso a la universidad

Modalidad	Nivel	Modalidad	Nivel
Diplomáticos y funcionarios internacionales	1	Tercio superior de colegios seleccionados B	6
Examen de admisión A	2	Bachillerato internacional sin convalidación	7
Deportistas calificados de alto rendimiento	3	Tercio superior de colegios seleccionados A	8
Examen de admisión B	4	Primeros puestos de secundaria	9
Centro preuniversitario	5		

Elaboración propia

El modelo construido con los datos del semestre 2016-1 se utilizó para predecir el resultado académico de los 958 ingresantes al Programa de Estudios Generales de la Universidad de Lima en el semestre 2016-2, agrupados en tres dominios de estudio, tal como se observa en la tabla 4.

Tabla 4
Distribución de ingresantes en el semestre 2016-2 por dominio de estudio

Dominio de estudio	Número de estudiantes
1. Carreras de Economía, Administración, Marketing, Negocios Internacionales, Contabilidad, Ingeniería Industrial e Ingeniería de Sistemas	646
2. Carreras de Psicología, Derecho y Comunicación	258
3. Carrera de Arquitectura	54
Total	**958**

Fuente: Universidad de Lima, Dirección Universitaria de Servicios Académicos y Registro (2016)
Elaboración propia

2.2 Modelo estadístico

En el modelo estadístico del análisis discriminante se definieron dos grupos a partir de la variable dependiente y se asignó una cantidad de

estudiantes para cada grupo en función de las medidas de 21 variables independientes $(X_1, ..., X_{21})$. Es decir, se trató de obtener para cada estudiante una puntuación Y, denominada *función discriminante*, de modo que fuera función lineal de $X_1, ..., X_{21}$. Esta función discriminante o combinación lineal de las 21 variables discriminó o separó lo más que se pudo los dos grupos de estudiantes, esto es, maximizó la varianza entre los grupos y minimizó la varianza dentro de los grupos (Marín, 2006).

El número de funciones discriminantes es determinado por $m = mín\{2-1,\ 21\} = 1$. Así, la función discriminante está dada de la siguiente manera:

$$Y = a_1 X_1 + \ \ldots \ + a_{21} X_{21} + a_0$$

Donde $a_1, a_2, \ldots, a_{21}$ son las ponderaciones de las variables independientes y a_0 es la constante.

La función Y se expresa como combinación lineal de $X_1, ..., X_{21}$ para que proporcione la mayor discriminación posible entre los grupos; en otras palabras, maximice la variabilidad entre los grupos y, al mismo tiempo, minimice la variabilidad dentro de estos (Peña, 2002).

2.3 Supuestos del modelo

El modelo estadístico del análisis multivariante depende de los siguientes supuestos:

- Las variables originales o discriminantes $(X_1, ..., X_{21})$ se deben distribuir como una normal multivariante. Si un conjunto de variables univariantes se distribuye como una normal multivariante, entonces la variable que es combinación lineal de ellas se distribuye como una normal univariante. Por esta razón, si alguna de las variables originales no se distribuye como una normal, entonces es seguro que todas las variables conjuntamente no se distribuirán como una normal multivariante.
- Las matrices de covarianzas de las variables discriminantes $(X_1, ..., X_{21})$ deben ser iguales en todos los grupos. Para comprobar esto, se puede usar la prueba M de Box. Dicha prueba tiene como hipótesis nula que las matrices de covarianzas son iguales. El test M de Box es sensible a la falta de normalidad multivariante, es decir, matrices iguales pueden aparecer como significativamente diferentes si no existe normalidad. Por otra parte, si las muestras son grandes, la prueba de Box pierde efectividad y, por tanto, es más fácil rechazar la hipótesis nula (Marín, 2006).

- Las variables discriminantes son no multicolineales entre sí, es decir, ninguna variable discriminante puede ser combinación lineal de otras variables discriminantes. Esto significa que son linealmente independientes.

En la práctica, el análisis discriminante es una técnica robusta y funciona bien, aunque los supuestos anteriores no se cumplan. A veces, estos supuestos son difíciles de probar. Algunos autores han demostrado que el análisis discriminante es una técnica robusta que tolera ciertas desviaciones de estos supuestos sin perder efectividad en las predicciones (Lachenbruch, 1975).

2.4 Selección de variables discriminantes

El método utilizado para la selección de las variables que conformarían la función discriminante fue el de inclusión por pasos. Con este método, las variables independientes son incorporadas paso a paso a la función discriminante tras evaluar su grado de contribución individual a la diferenciación de los grupos. El método se inicia seleccionando la mejor variable independiente (es decir, la que más discrimina los grupos), pero solo si esta cumple el criterio de entrada. En concreto, una variable independiente se incluye en la función discriminante si el valor del estadístico F de Fisher-Snedecor es mayor de 3,84 (al 5 % de significación). Luego, se selecciona la variable independiente que cumple el criterio de entrada y es la que más contribuye a que la función discriminante diferencie los grupos. Cada vez que se incorpora una nueva variable al modelo, las variables previamente seleccionadas son evaluadas otra vez para determinar si cumplen el criterio de salida; de este modo, una variable independiente es eliminada de la función discriminante si el valor del estadístico F es menor de 2,71 (al 10 % de significación). Si alguna variable de las ya seleccionadas cumple el criterio de salida, es eliminada del modelo.

Una vez que se aplicó el criterio de significación anterior, se analizó la multicolinealidad entre las variables independientes. En este sentido, una variable solo pasa a formar parte de la función discriminante si su *nivel de tolerancia*[1] es mayor que el nivel especificado y si, además, su inclusión en la función no ocasiona que alguna de las variables

1 La tolerancia de una variable independiente es la proporción de varianza de esa variable que no está relacionada con el resto de las variables independientes.

previamente seleccionadas alcance un nivel de tolerancia por debajo del nivel establecido.

Las variables seleccionadas que fueron incluidas en la función discriminante para cada dominio de estudio sobre la base de los criterios de significación y multicolinealidad se presentan en la tabla 5, según el orden de selección.

Tabla 5
Variables seleccionadas para cada dominio de estudio utilizando el método de inclusión por pasos

Dominio 1	Dominio 2	Dominio 3
X_3: PPA del parcial	X_3: PPA del parcial	X_{11}: nota del trabajo 1 de DPS
X_{13}: porcentaje de inasistencias a MI	X_{16}: porcentaje de inasistencias a GRN	X_{18}: nota del trabajo 1 de D1
X_2: cursos desaprobados en el parcial	X_2: cursos desaprobados en el parcial	X_2: cursos desaprobados en el parcial
X_8: nota del trabajo 1 de LC1	X_{21}: nota del PA1 de FM	
X_{14}: nota del trabajo 1 de MI	X_{10}: porcentaje de inasistencias a DPS	
X_{16}: porcentaje de inasistencias a GRN	X_{12}: nota del parcial de MI	
X_{15}: nota del parcial de GRN	X_{20}: nota de práctica calificada 1 de FM	
X_4: nota del parcial de LC1	X_{11}: nota del trabajo 1 de DPS	
X_{19}: % inasistencias a MB		

Elaboración propia

2.5 Regla de discriminación: funciones de clasificación de Fisher

Otra forma de utilizar la función discriminante consiste en construir funciones discriminantes para cada grupo. Estas se denominan funciones de clasificación de Fisher y tienen la siguiente forma (Peña, 2002):

$$F_0 = a_{0,1}X_1 + \ldots + a_{0,21}X_{21} - C_0$$
$$F_1 = a_{1,1}X_1 + \ldots + a_{1,21}X_{21} - C_1$$

Para cada observación, se calculó el valor de la función de clasificación de Fisher en cada uno de los grupos y se clasificó la observación en el grupo con el valor más grande para esa función de clasificación, es decir:

$$\begin{cases} \text{Si } F_0 > F_1 \text{ el ingresante pertenece al grupo } 0. \\ \text{Si } F_0 < F_1 \text{ el ingresante pertenece al grupo } 1. \end{cases}$$

Existen otras reglas de clasificación equivalentes a la regla de clasificación de Fisher, tales como la distancia de Mahalanobis y la de probabilidades de Bayes, solo es asunto de cómo se prefiera ver la solución para el problema de discriminación (Dallas, 2000).

En este trabajo se utilizó la función discriminante de Fisher para la etapa de entrenamiento (en la que se estimaron los modelos con los datos de los ingresantes del semestre 2016-1 con sus grupos de pertenencia conocidos) y las funciones de clasificación de Fisher para la etapa de clasificación (en la que se usaron los modelos estimados para clasificar a los ingresantes del 2016-2). Los coeficientes de estas funciones fueron estimados para los tres dominios de estudio. Los resultados se presentan en la siguiente sección.

3. RESULTADOS Y DISCUSIÓN

Se desarrollaron tres modelos de predicción, uno por cada dominio de estudio, que consistieron en las funciones de clasificación de Fisher. Con estos modelos, se identificaron a los estudiantes que aprobarían a lo sumo dos de los cinco cursos en el semestre 2016-2.

Se utilizaron los programas estadísticos SPSS y R para procesar los datos. Ambos programas arrojaron los mismos resultados, pero con reportes de diferentes medidas de rendimiento. Entre los indicadores más importantes para determinar el rendimiento de los modelos de predicción, se consideró la proporción de estudiantes clasificados correctamente, medida conocida como *exactitud*. A continuación, se presentaron la función discriminante estimada con las variables significativas, las funciones de clasificación de Fisher y la matriz de confusión con los indicadores de rendimiento del modelo de predicción.

3.1 Dominio 1

Los coeficientes de la función discriminante estimada se muestran en la tabla 6. Esta función solo incluye las variables seleccionadas mediante el método de inclusión por pasos y la constante.

Tabla 6
Coeficientes estimados de la función discriminante del dominio 1

Variable independiente	Función discriminante
X_2: cursos desaprobados en el parcial	0,428
X_3: PPA del parcial	–0,169
X_4: nota del parcial de LC1	0,055
X_8: nota del trabajo 1 de LC1	–0,096
X_{13}: porcentaje de inasistencias a MI	0,050
X_{14}: nota del trabajo 1 de MI	–0,051
X_{15}: nota del parcial de GRN	0,047
X_{16}: porcentaje de inasistencias a GRN	0,031
X_{19}: porcentaje de inasistencias a MB	0,032
a_2: constante	2,538

Elaboración propia

De la tabla 6, la función discriminante está dada por

$$Y = 0,428X_2 - 0,169X_3 + 0,055X_4 - 0,096X_8 + 0,05X_{13} - 0,051X_{14} + 0,047X_{15} + 0,031X_{16} + 0,032X_{19} + 2,538$$

Las funciones de clasificación de Fisher estimadas para cada grupo son las siguientes:

Grupo 0: hasta 2 cursos aprobados

$$F_1 = 15,724X_2 + 6,558X_3 - 0,013X_4 + 1,864X_8 + 0,81X_{13} + 0,534X_{14} + 0,346X_{15} + 0,425X_{16} + 0,581X_{19} - 78,029$$

Grupo 1: de 3 a 5 cursos aprobados

$$F_2 = 14,419X_2 + 7,074X_3 - 0,18X_4 + 2,158X_8 + 0,657X_{13} + 0,688X_{14} + 0,204X_{15} + 0,331X_{16} + 0,485X_{19} - 78,838$$

Según las funciones de Fisher obtenidas, se observa que las variables que más contribuyen a los valores de dichas funciones son X_2 y X_3, debido a la magnitud de sus coeficientes.

En la tabla 7 se presenta la matriz de confusión para el dominio 1. En esta matriz se muestran los resultados de predicción para los ingresantes matriculados en el semestre 2016-2 frente a los resultados reales.

Tabla 7
Matriz de confusión para el dominio 1

Resultados de predicción	Resultados reales		Total
	Hasta 2 cursos aprobados	De 3 a 5 cursos aprobados	
Hasta 2 cursos aprobados	42	21	63
De 3 a 5 cursos aprobados	26	557	583
Total	68	578	646

Elaboración propia

De la tabla 7, se obtuvieron las medidas de rendimiento del modelo de predicción para el dominio 1. Las medidas de rendimiento más importantes son las siguientes:

1. *Exactitud*

$$Accuracy = \frac{42 + 557}{646} = 0,9272$$

Esta medida indica que el 92,72 % del total de ingresantes matriculados fueron clasificados correctamente; es decir, el modelo de predicción fue efectivo en el 92,72 % de los casos estudiados.

2. *Sensitivity* (sensibilidad)

$$Sensitivity = \frac{42}{68} = 0,6177$$

Esta medida indica que el 61,77 % de los ingresantes matriculados que aprobaron hasta dos cursos fueron clasificados correctamente; es decir, el modelo de predicción en el grupo de ingresantes que aprobaron hasta dos cursos fue efectivo en el 61,77 % de los casos estudiados.

3. *Specificity* (especificidad)

$$Specificity = \frac{557}{578} = 0,9637$$

Esta medida indica que el 96,37 % de los ingresantes matriculados que aprobaron de tres a cinco cursos fueron clasificados correctamente; es decir, el modelo de predicción en el grupo de ingresantes

que aprobaron de tres a cinco cursos fue efectivo en el 96,37 % de los casos estudiados.

3.2 Dominio 2

Los coeficientes de la función discriminante estimada se presentan en la tabla 8. Esta función solo incluye las variables seleccionadas mediante el método de inclusión por pasos y la constante.

Tabla 8
Coeficientes estimados de la función discriminante del dominio 2

Variable independiente	Función discriminante
X_2: cursos desaprobados en el parcial	0,398
X_3: PPA del parcial	–0,152
X_{10}: porcentaje de inasistencias a DPS	0,080
X_{11}: nota del trabajo 1 de DPS	–0,023
X_{12}: nota del parcial de MI	0,075
X_{16}: porcentaje de inasistencias a GRN	0,085
X_{20}: nota de práctica calificada 1 de FM	0,045
X_{21}: nota del práctica de aula 1 FM	–0,085
a_0: constante	1,275

Elaboración propia

De la tabla 8, la función discriminante está dada por

$$Y = 0,398X_2 - 0,152X_3 + 0,08X_{10} - 0,023X_{11} + 0,075X_{12} + 0,085X_{16} + 0,045X_{20} - 0,085X_{21} + 1,275$$

Las funciones de clasificación de Fisher estimadas para cada grupo son las siguientes:

Grupo 0: hasta 2 cursos aprobados

$$F_1 = 16,35X_2 + 5,621X_3 + 0,6X_{10} + 0,204X_{11} + 1,192X_{12} + 1,12X_{16} - 0,031X_{20} + 0,955X_{21} - 72,628$$

Grupo 1: de 3 a 5 cursos aprobados

$$F_2 = 14,801X_2 + 6,212X_3 + 0,288X_{10} + 0,292X_{11} + 0,9X_{12} + 0,791X_{16} - 0,205X_{20} + 1,287X_{21} - 67,692$$

Según las funciones de Fisher obtenidas, se observa que las variables que más contribuyen a los valores de dichas funciones son X_2 y X_3, debido a la magnitud de sus coeficientes.

En la tabla 9 se muestra la matriz de confusión para el dominio 2. En esta matriz se presentan los resultados de predicción para los ingresantes matriculados en el semestre 2016-2 frente a los resultados reales.

Tabla 9
Matriz de confusión para el dominio 2

Resultados de predicción	**Resultados reales**		**Total**
	Hasta 2 cursos aprobados	**De 3 a 5 cursos aprobados**	
Hasta 2 cursos aprobados	13	7	20
De 3 a 5 cursos aprobados	13	225	238
Total	26	232	258

Elaboración propia

De la tabla 9, se obtuvieron las medidas de rendimiento del modelo de predicción para el dominio 2. Las medidas de rendimiento más importantes son las siguientes:

1. *Accuracy* (exactitud)

$$Accuracy = \frac{13 + 225}{258} = 0{,}9225$$

Esta medida indica que el 92,25 % del total de ingresantes matriculados fueron clasificados correctamente; es decir, el modelo de predicción fue efectivo en el 92,25 % de los casos estudiados.

2. *Sensitivity* (sensibilidad)

$$Sensitivity = \frac{13}{26} = 0{,}5$$

Esta medida indica que el 50 % de los ingresantes matriculados que aprobaron hasta dos cursos fueron clasificados correctamente; es decir, el modelo de predicción en el grupo de estudiantes que aprobaron hasta dos cursos fue efectivo en el 50 % de los casos estudiados.

3. *Specificity* (especificidad)

$$Specificity = \frac{225}{232} = 0,9698$$

Esta medida indica que el 96,98 % de los ingresantes matriculados que aprobaron de tres a cinco cursos fueron clasificados correctamente; es decir, el modelo de predicción en el grupo de ingresantes que aprobaron de tres a cinco cursos fue efectivo en el 96,98 % de los casos estudiados.

3.3 Dominio 3

Los coeficientes de la función discriminante estimada se presentan en la tabla 10. Esta función solo incluye las variables seleccionadas mediante el método de inclusión por pasos y la constante.

Tabla 10
Coeficientes estimados de la función discriminante del dominio 3

Variable independiente	Función discriminante
X_{11}: nota del trabajo 1 de DPS	0,213
X_{18}: nota del trabajo 1 de D1	0,221
a_0: Constante	–6,365

Elaboración propia

De la tabla 10, la función discriminante está dada por

$$Y = 0,213X_{11} + 0,221X_{18} - 6,365$$

Las funciones de clasificación de Fisher estimadas para cada grupo son las siguientes:

Grupo 0: hasta 2 cursos aprobados

$$F_1 = 0,473X_{11} + 1,352X_{18} - 10,918$$

Grupo 1: de 3 a 5 cursos aprobados

$$F_2 = 1,072X_{11} + 1,974X_{18} - 22,775$$

Según las funciones de Fisher obtenidas, se observa que la variable que más influye en los valores de dichas funciones es X_{18}, debido a la magnitud de su coeficiente.

En la tabla 11 se presenta la matriz de confusión para el dominio 3. En esta matriz se muestran los resultados de predicción para los ingresantes matriculados en el semestre 2016-2 frente a los resultados reales.

Tabla 11
Matriz de confusión para el dominio 3

Resultados de predicción	**Resultados reales**		**Total**
	Hasta 2 cursos aprobados	**De 3 a 5 cursos aprobados**	
Hasta 2 cursos aprobados	4	3	7
De 3 a 5 cursos aprobados	7	40	47
Total	11	43	54

Elaboración propia

De la tabla 11, se obtuvieron las medidas de rendimiento del modelo de predicción para el dominio 3. Las medidas de rendimiento más importantes son las siguientes:

1. *Exactitud*

$$Accuracy = \frac{4+40}{54} = 0,8148$$

Esta medida indica que el 81,48 % del total de ingresantes matriculados fueron clasificados correctamente; es decir, el modelo de predicción fue efectivo en el 81,48 % de los casos estudiados.

2. *Sensitivity* (sensibilidad)

$$Sensitivity = \frac{4}{11} = 0,3636$$

Esta medida indica que el 36,36 % de los ingresantes matriculados que aprobaron hasta dos cursos fueron clasificados correctamente; es decir, el modelo de predicción en el grupo de ingresantes que aprobaron hasta dos cursos fue efectivo en el 36,36 % de los casos estudiados.

3. *Specificity* (especificidad)

$$Specificity = \frac{40}{43} = 0,9302$$

Esta medida indica que el 93,02 % de los ingresantes matriculados que aprobaron de tres a cinco cursos fueron clasificados correctamente; es decir, el modelo de predicción en el grupo de ingresantes que aprobaron de tres a cinco cursos fue efectivo en el 93,02 % de los casos estudiados.

3.4 Análisis de la curva ROC

La curva ROC es un gráfico en el que se observan los pares $(1-E; S)$ resultantes de la variación continua de los puntos de corte en todo el rango de resultados observados. El eje X representa a $(1-E)$, donde E es la especificidad; y el eje Y representa la sensibilidad (S).

Cada punto de la curva está dado por un par $(1-E; S)$ que corresponde a un nivel de decisión determinado. Un modelo con discriminación perfecta, sin solapamiento de resultados en los dos grupos, muestra una curva ROC que pasa por la esquina superior izquierda, donde E y S toman valores máximos $(E = S = 1)$. Por otro lado, un modelo sin discriminación, con igual distribución de resultados en los dos grupos, genera una línea diagonal de 45º, desde la esquina inferior izquierda hasta la superior derecha. La mayoría de las curvas ROC caen por encima de la diagonal.

El área bajo la curva (ABC) ROC es una medida global de la exactitud del modelo. Se define como la probabilidad de clasificar correctamente un par de ingresantes crítico (aprobaron a lo más dos cursos) y un par no crítico (aprobaron de tres a cinco cursos) seleccionados al azar de la población. El ABC se interpreta de la siguiente manera: valores entre 0,5 y 0,7 indican baja exactitud; entre 0,7 y 0,9 señalan exactitud moderada y pueden ser útiles para algunos propósitos; y un valor mayor de 0,9 indica exactitud alta.

La capacidad de discriminación del modelo puede evaluarse estimando el intervalo de confianza del ABC ROC. Si el intervalo no incluye el valor 0,5, la prueba es capaz de discernir entre críticos y no críticos (Burgueño, García Bastos y González Buitrago, 1995).

Los resultados de las áreas ABC ROC para los tres dominios de estudio se presentan en la tabla 12.

Tabla 12
ABC ROC de los tres dominios de estudio

Dominio	Área	Error estándar[a]	Valor p[b]	Intervalo de confianza asintótico al 95 %	
				Límite inferior	Límite superior
1	0,826	0,028	0,000	0,770	0,881
2	0,823	0,061	0,000	0,703	0,944
3	0,842	0,074	0,004	0,697	0,987

Nota:
[a] Bajo el supuesto no paramétrico
[b] Hipótesis nula: área verdadera = 0,5
Elaboración propia

La curva ROC para cada dominio de estudio se muestra en la figura 1.

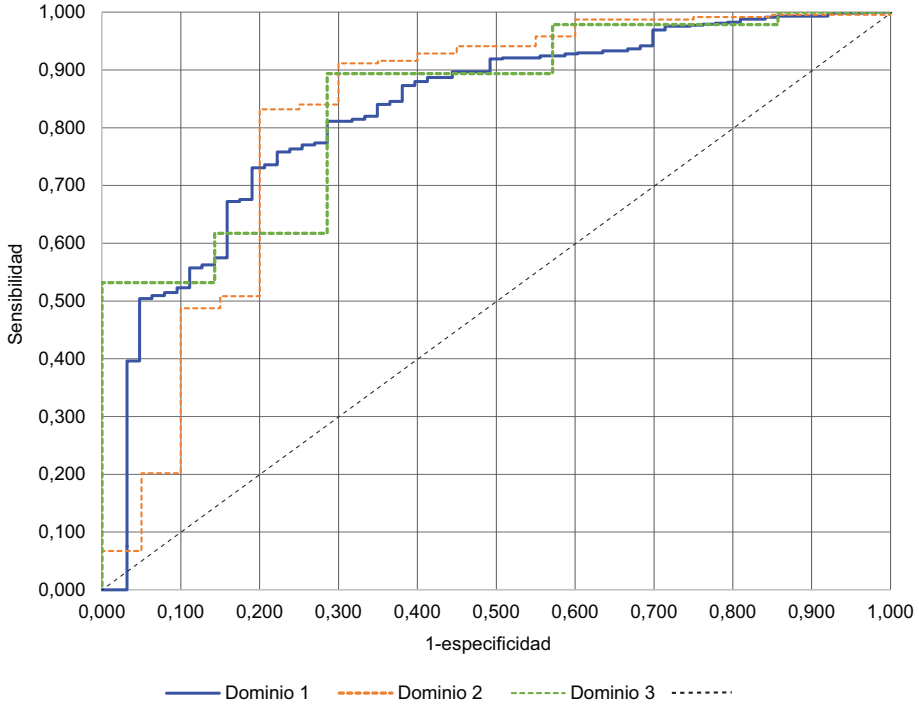

Figura 1. Curvas ROC por dominio de estudio
Elaboración propia

De acuerdo con las ABC ROC de la tabla 12 y la figura 1, la exactitud del modelo en los tres dominios de estudio es moderada. Esto significa que tales modelos son útiles para la detección temprana de ingresantes críticos y no críticos.

4. CONCLUSIONES

La técnica empleada en este trabajo fue el análisis discriminante. Esta herramienta estadística permitió construir los modelos para clasificar a los ingresantes del Programa de Estudios Generales de la Universidad de Lima del semestre 2016-1 y luego predecir, con los mismos modelos, el rendimiento académico de los ingresantes del 2016-2.

El uso del análisis discriminante, para predecir el rendimiento académico de los ingresantes del Programa de Estudios Generales de la Universidad de Lima, constituye una herramienta útil por las siguientes razones:

- Los ingresantes matriculados clasificados correctamente, según los modelos de predicción, resultaron por encima del 92 % en los dominios 1 y 2, y cerca del 80 % en el dominio 3. En resumen, la efectividad fue alta.
- Los ingresantes que aprobaron hasta dos cursos fueron clasificados correctamente, según los modelos de predicción, en el 62 % y el 50 % de los casos en los dominios 1 y 2, respectivamente. En dichos dominios las muestras fueron relativamente grandes, en otras palabras, la sensibilidad fue aceptable. En el dominio 3 solo se alcanzó una sensibilidad del 27 %, debido a que la muestra fue muy pequeña.
- Los ingresantes que aprobaron de tres a cinco cursos fueron clasificados correctamente, según los modelos de predicción, en el 96 %, el 96 % y el 93 % de los casos en los dominios 1, 2 y 3, respectivamente, lo que indica que la especificidad fue alta.

El análisis de la curva ROC evaluó la exactitud de los modelos utilizados en los tres dominios de estudio, exactitud que fue moderada en los tres. Este resultado permite identificar con cierta seguridad y tempranamente a los ingresantes críticos y los no críticos.

Por lo expuesto, la técnica del análisis discriminante realizó una buena predicción con alto porcentaje de efectividad y especificidad en los tres dominios, con sensibilidad aceptable en los dominios 1 y 2, pero con

baja sensibilidad en el dominio 3. Estos niveles de sensibilidad se pueden mejorar con el uso de otras técnicas de predicción, como redes neuronales.

REFERENCIAS

Bloom, B. S. (1976). *Características humanas y aprendizaje escolar*. Bogotá: Voluntad.

Burgueño, M.-J., García Bastos, J., y González Buitrago, J. (1995). Las curvas ROC en la evaluación de las pruebas diagnósticas. *Medicina clínica*, *104*(17), 661-670.

Carvajal, P., Trejos, A., y Soto, J. (2004). Aplicación del análisis discriminante para explorar la relación entre el examen de ICFES y el rendimiento en Álgebra Lineal de los estudiantes de Ingeniería de la UTP en el periodo 2001-2003. *Scientia et Technica*, *2*(25), 191-196.

Dallas, E. J. (2000). *Métodos multivariados aplicados al análisis de datos*. Ciudad de México: Thomson.

Di-Gresia, L., Porto, A., y Ripani, L. (noviembre del 2002). *Rendimiento de los estudiantes de las universidades públicas argentinas* (DT. N.° 45). Recuperado de http://sedici.unlp.edu.ar/handle/10915/3476

Ezequiel, J. (2005). *Análisis multivariante aplicado*. Madrid: Thomson-Paraninfo.

Ibarra, M., y Michalus, J. (2010). Análisis del rendimiento académico mediante un modelo Logit. *Ingeniería Industrial*, *9*(2), 47-56.

Lachenbruch, P. A. (1975). Zero-Mean Difference Discrimination and the Absolute Linear Discriminant Function. *Biometrika*, *62*(2), 397-401.

Marín, M.-A. (2006). *Análisis multivariante*. Universidad Carlos III de Madrid.

Peña, D. (2002). *Análisis de datos multivariantes*. Madrid: McGraw-Hill.

Reyes Tejada, Y. (2003). *Relación entre el rendimiento académico, la ansiedad ante los exámenes, los rasgos de personalidad, el autoconcepto y la asertividad en estudiantes del primer año de*

Psicología de la UNMSM (tesis de licenciatura para optar por el título de psicólogo). Universidad Nacional Mayor de San Marcos.

Universidad de Lima, Dirección Universitaria de Servicios Académicos y Registro. (2016). Distribución de ingresantes en el semestre 2016-1 por dominio de estudio [documento interno].

Universidad de Lima, Dirección Universitaria de Servicios Académicos y Registro. (2016). Distribución de ingresantes en el semestre 2016-2 por dominio de estudio [documento interno].