Anales Científicos, 77 (1): 8-16 (2016)
ISSN 0255-0407 (Versión impresa)
ISSN 2519-7398 (Versión electrónica)

DOI: http://dx.doi.org/10.21704/ac.v77i1.474

Presentado:19/12/2014
Aceptado:11/03/2015

PREDICCIÓN DEL RENDIMIENTO EN EL EXAMEN DE ADMISIÓN A LA UNALM UTILIZANDO LAS TÉCNICAS DE ANÁLISIS DISCRIMINANTE LINEAL Y ANÁLISIS DISCRIMINANTE CON ALGORITMOS GENÉTICOS

PREDICTION OF PERFORMANCE INTO ENTRANCE EXAMINATION TO UNALM USING LINEAR ANALYSIS DISCRIMINANT AND DISCRIMINANT ANALYSIS WITH GENETIC ALGORITHMS

Joao M. Rado H[1]., [2]Jesús W. Salinas F.y [3]Fernando R. Rosas V.

Resumen

El objetivo de la investigación fue probar la hipótesis que la tasa de error de clasificación utilizando el análisis discriminante con algoritmos genéticos es menor a la que se obtiene con el análisis discriminante lineal de Fisher. La aplicación se efectuó en la predicción del rendimiento en el examen de admisión de la Universidad Nacional Agraria La Molina de los postulantes cuya preparación se realizó en su Centro de Estudios Preuniversitarios. En la técnica de algoritmos genéticos se empleó el método de selección, cruce y mutación que permitió realizar la búsqueda de funciones discriminantes con error mínimo. Los resultados del estudio indican que el análisis discriminante con algoritmos genéticos proporcionó una función discriminante más eficiente que la proporcionada por Fisher.

Palabras clave: Análisis Discriminante, Algoritmos Genéticos, Optimización.

Abstract

The aim of the research was to test the hypothesis that the error rate of classification using discriminant analysis with genetic algorithms is lower than obtained with the Fisher linear discriminant analysis. The study was made in predicting performance in the entrance examination of the Universidad NacionalAgraria La Molina of applicants whose preparation was conducted in the Preparatory School of the UNALM. In the technique of genetic algorithms your method of selection, crossover and mutation allowing search discriminant function with minimal error was used. The results indicate that the discriminant analysis with genetic algorithms provided a more efficient discriminant function that provided by Fisher.

Keys words: Discriminant Analysis, Genetic Algorithms, Optimization.

1. Introducción

En el área de Estadística se han venido desarrollando técnicas de análisis multivariadocon fines de clasificar objetos o individuos. Entre éstas técnicas se encuentran el análisis clúster, el análisis de regresión logística, el análisis discriminante entre otras.

El Análisis Discriminante Lineal propuesto por Fisher (1936), tiene como objetivo determinar las variables que explican mejor la pertenencia de un individuo a un determinado grupo, y además estima una función discriminante que permite clasificarlo en uno de los grupos existentes.

A partir de este modelo sehan presentado notables avances, desde el Análisis Discriminante Flexible (Hastie, T., Tibshirani, R. y Buja, A, 1994.) y Análisis Discriminante Penalizado (Hastie, T., Tibshirani, R. y Buja, A, 1995.), hasta los más recientes basados en remuestreo (Breiman, 1998) y las redes neuronales artificiales (López, M, et al., 2007), que buscan reducir al máximo la tasa de error de clasificación. Uno de éstos últimos es el Análisis discriminante con algoritmos genéticos, que utiliza resultados del análisis discriminante lineal y operadores genéticos para estimar una o varias funciones discriminantes.

El objetivo de la investigación es comparar la eficiencia del análisis discriminante con algoritmos genéticos respecto al análisis discriminante lineal de Fisher a través de la tasa de error de clasificación. Para ello, se postula la hipótesis que la tasa de error de clasificación utilizando el análisis discriminante con algoritmos genéticos es menor al que se obtiene con el análisis discriminante lineal de Fisher.

La hipótesis de investigación se sometió a prueba en una aplicación para predecir el perfil de rendimiento en el examen de admisión de la Universidad Nacional Agraria La Molina (UNALM) de los postulantes que ingresaron o no a la universidady cuya preparación se realizó en su Centro de Estudios Preuniversitarios (CEPRE_UNALM).

[1]Departamento Académico de Estadística e Informática de la UNALM. E-mail: jrado@lamolina.edu.pe
[2]Departamento Académico de Estadística e Informática de la UNALM. E-mail: jsalinas@lamolina.edu.pe
[3]Departamento Académico de Estadística e Informática de la UNALM. E-mail: frosas@lamolina.edu.pe

La data correspondió a los resultados de las pruebas de admisión de la UNALM de los concursos de admisión del período 2009 – 2013.

2. Materiales y métodos

Metodología de la Investigación

Tipo de la Investigación

El tipo de investigación es de carácter descriptivo y correlacional/causal en la aplicación que se realizó en ambas técnicas discriminantes en el concurso de admisión(2009-2013), debido a la presencia de una variable dependiente (Y) de naturaleza categórica (dicotómica) y la de seis variables independientes (X_1, X_2, X_3, X_4, X_5, X_6) de naturaleza cuantitativa.

Diseño de la investigación

El diseño de la investigación fue de tipo no experimental-transversal, ya que se obtiene un conjunto de datos provenientes de los resultados de los exámenes de admisión comprendidos entre los años 2009 al 2013.

Instrumento de colecta de datos

El instrumento empleado en la obtención de los datos requeridos para la investigación fue el examen de admisión elaborado por el Comité Permanente de Admisión. Este instrumento tiene un tiempo de aplicación de aproximadamente tres horas, 100 preguntas con cinco alternativas, donde sólo hay una respuesta correcta. Las preguntas están distribuidas en nueve cursos de la siguiente forma: Razonamiento Matemático (14), Razonamiento Verbal (20), Aritmética (8), Algebra (6), Geometría (6), Trigonometría (4), Física (14), Química (14) y Biología (14). Cada pregunta bien contestada tiene un valor de 1.00 punto, sin contestar 0.00 y mal contestada – 0.25.

Formulación de las hipótesis

Las hipótesis que corresponden al presente trabajo de investigación son las siguientes:

1. El análisis discriminante con algoritmos genéticos proporciona una tasa de error de clasificación menor para predecir el rendimiento en el examen de admisión de la UNALM cuya preparación se realizó en el CEPRE-UNALM que la proporcionada por el análisis discriminante lineal de Fisher.
2. Los nueve cursos que se evalúan en la prueba de Admisión de la UNALM tienen capacidad discriminante en el perfil del rendimiento de los postulantes cuya preparación se realizó en el CEPRE-UNALM.

Identificación de las variables

Y = Rendimiento en el examen del Concurso de Admisión de la UNALM de los postulantes que ingresaron o no a la universidad y cuya preparación se realizó en el CEPRE-UNALM.

Esta variable es considerada como la variable dependientey tiene dos categorías:

- No ingresó a la universidad.
- Si ingresó a la universidad.

Las variables independientes o predictoraslo constituyen los nueve cursos que se imparten en el CEPRE-UNALM:

X_1= puntaje obtenido en Razonamiento Matemático.
X_2= puntaje obtenido en Razonamiento Verbal.
X_3= puntaje obtenido en Matemática (Álgebra, Aritmética, Geometría y Trigonometría).
X_4= puntaje obtenido en Física.
X_5= puntaje obtenido en Química.
X_6= puntaje obtenido en Biología.

Definiciones operacionales

En la aplicación del análisis discriminante y análisis discriminante con algoritmos genéticos la variable dependiente (Y) es de naturaleza categórica (dicotómica) y las variables independientes X_1, X_2, X_3, X_4, X_5, X_6 son de naturaleza cuantitativa medidas en una escala vigesimal.

Población y muestra

La población son todos los postulantes al Examen Ordinario de Admisión de la UNALM que realizaron sus estudios preuniversitarios en el CEPRE-UNALM.

En la investigación se trabajó con una muestra de 3840 postulantes al Examen Ordinario de Admisión de la UNALM (2009-2013) que realizaron sus estudios preuniversitarios en el CEPRE-UNALM; de los cuales 590 fueron ingresantes y 3250 no ingresantes.

Metodología aplicada

Los pasos que se realizaron para llevar a cabo este trabajo se detallan a continuación:

1. Análisis estadístico univariado

   Para cada variable independiente, se obtuvo la media y desviación estándar por grupo.

2. Análisis estadístico bivariado

   Se utilizó el gráfico de dispersión entre las variables independientes por cada grupo.

3. Análisis discriminante lineal

3.1 Verificación de Supuestos

3.2 Análisis de las variables explicativas

3.3 Función discriminante lineal de Fisher

3.4 Validación cruzada

4. Análisis discriminante con algoritmos genéticos

4.1 Generación de la población inicial

4.2 Método de la ruleta

   Se calculó la función de aptitud en base a la tasa promedio de clasificación correcta, de tal forma que se vean beneficiadas las funciones con menor error.

4.3 Funciones obtenidas mediante cruce y mutación

4.4 Función discriminante óptima

4.5 Validación cruzada en Algoritmos Genéticos

5. Comparación de resultados del Análisis discriminante lineal y Análisis discriminante con algoritmos genéticos.

3. Resultados y discusión

Antes de efectuar el análisis de clasificación se realizó una limpieza y consistencia de datos.

Análisis estadístico univariado

Tabla 1. Estadísticos de grupo.

INGRESO		Media	Desv. típ. No ponderados	N válido (según lista) ponderados No	Si
No Ingreso	RM	88.143	350.320	3226	3.226.000
	RV	89.110	337.889	3226	3.226.000
	MAT	61.921	416.514	3226	3.226.000
	FIS	49.235	389.772	3226	3.226.000
	QUI	83.511	507.840	3226	3.226.000
	BIO	49.860	395.024	3226	3.226.000
Ingreso	RM	129.012	297.240	573	573.000
	RV	114.629	317.956	573	573.000
	MAT	128.660	299.597	573	573.000
	FIS	108.508	325.120	573	573.000
	QUI	143.593	308.542	573	573.000
	BIO	94.624	389.469	573	573.000
Total	RM	94.308	372.710	3799	3.799.000
	RV	92.959	347.150	3799	3.799.000
	MAT	71.987	466.790	3799	3.799.000
	FIS	58.175	435.811	3799	3.799.000
	QUI	92.573	528.751	3799	3.799.000
	BIO	56.612	425.463	3799	3.799.000

Fuente: Elaboración Propia

En la Tabla 1 se puede apreciar que los postulantes que ingresaron a la UNALM poseen un mejor rendimiento promedio en todos los cursos frente a los que no ingresaron. Respecto a la variabilidad (desviación estándar), se puede observar que el grupo de no ingresantes posee mayor variabilidad en todos los cursos frente a los que ingresaron.

Análisis estadístico bivariado

En las Figuras1 y 2 se puede observar que los rendimientos se encuentran correlacionados en los 6 cursos para los ingresantes y no ingresantes. Este resultado da un indicio de que se encuentrae problemas de multicolinealidad.

Análisis discriminante lineal

Verificación de Supuestos

Se consideró necesario realizar la verificación de supuestos con la finalidad de que los resultados de la clasificación no se vean afectados.

Normalidad

Se realizó un análisis univariado preliminar de tipo descriptivo para verificar la normalidad utilizando el histograma, posteriormente fue corroborado con la prueba de normalidad multivariante.

En las Figuras1 y 2se presentaron también los gráficos de Histogramas de rendimientos de los 6 cursos para los ingresantes y no ingresantes. Se puede observar que para el grupo de ingresantes los cursos que aparentemente no cumplen con un ajuste de normalidad son Razonamiento Matemático, Razonamiento Verbal, Química y Biología. Mientras que para el grupo de no ingresantes los cursos de Matemática y Física son los que aparentemente no tienen un buen ajuste de normalidad.

Se realizó la prueba de normalidad multivariada de ShapiroWilk en R, para los ingresantes y no ingresantes. La función usada fue mshapiro.test del paquete mvnormtest.

Tanto para los ingresantes (P-valor=5.354e-08) como no ingresantes (P-valor=0.004489) no se cumplió la normalidad multivariada a un nivel de significación del 1%.

Homogeneidad de Matrices Varianza-Covarianza

Se realizó el análisis univariado de homogeneidad de varianzas mediante la prueba de Levene, posteriormente se hizo la prueba M de Box.

En la Tabla 2 se puede apreciar que existe homogeneidad de varianzas entre los grupos de ingresantes y no ingresantes para los rendimientos de los cursos de Razonamiento Verbal y Biología a un nivel de significación del 1%. Mientras que para los rendimientos en Razonamiento Matemático, Matemática, Física y Química (cada uno con P-valor=0.000) no existe homogeneidad de varianzas.

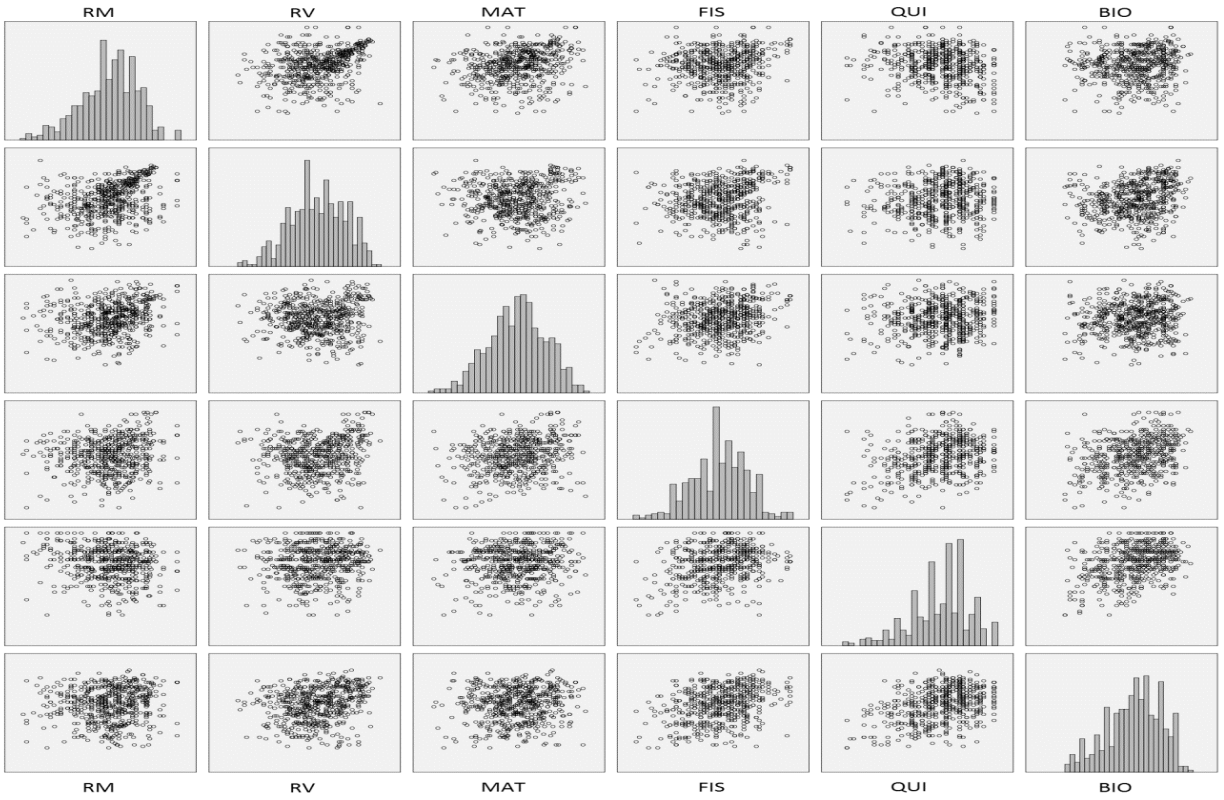

Figura 1. Gráfico de Dispersión para Ingresantes.
Fuente: Elaboración Propia

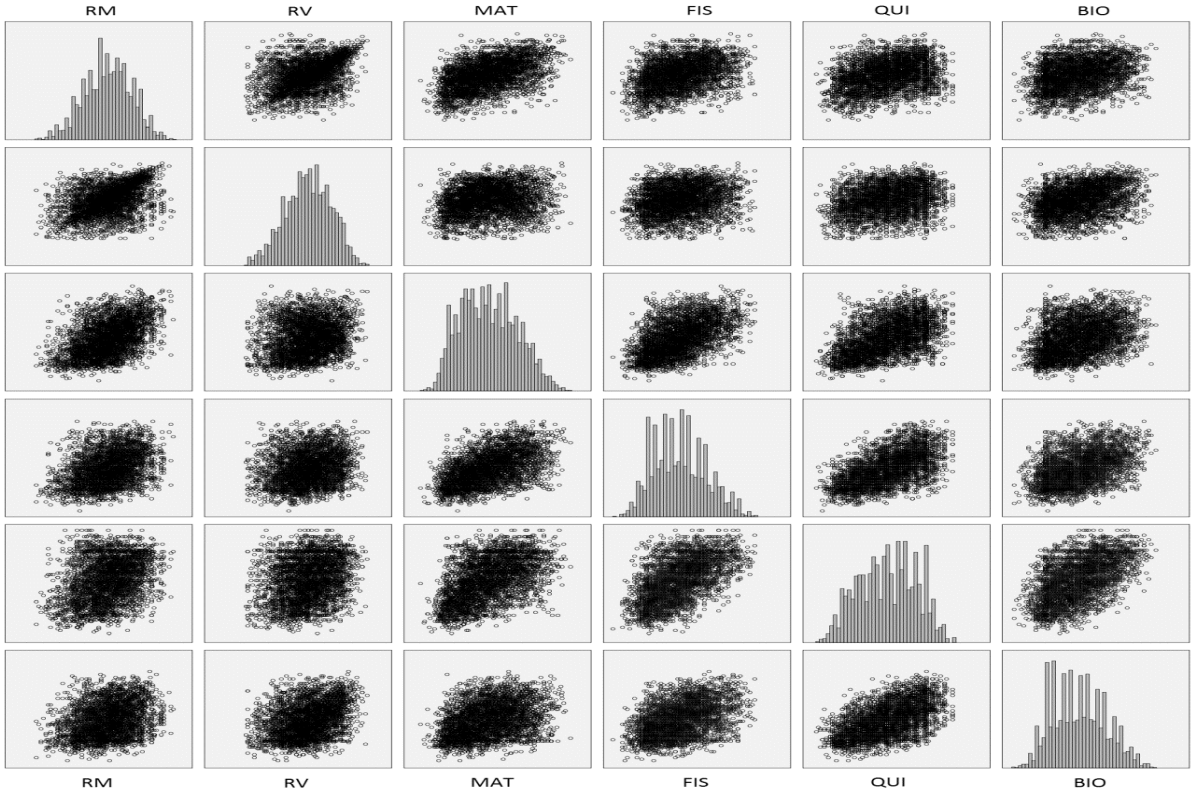

Figura 2. Gráfico de Dispersión para No Ingresantes.
Fuente: Elaboración Propia

Tabla 2. Prueba de Homogeneidad de Varianzas de Levene.

	Estadístico de Levene	gl1	gl2	Sig.
RM	29.039	1	3797	0
RV	2.159	1	3797	0.142
MAT	111.344	1	3797	0
FIS	38.918	1	3797	0
QUI	232.096	1	3797	0
BIO	1.465	1	3797	0.226

Fuente: Elaboración Propia

En la Tabla 3 se puede observar que la prueba M de Box resultó significativa al 1%, esto quiere decir que no se cumplió el supuesto de Homogeneidad de Matrices Varianza Covarianza.

Tabla 3. Prueba M de box.

M de Box		347.186
F	Aprox.	16.47
	gl1	21
	gl2	3778544.94
	Sig.	0

Fuente: Elaboración Propia

Multicolinealidad

Como una mejora a la posible multicolinealidad se aplicó un análisis factorial a los datos. Posteriormente, a las puntuaciones factoriales obtenidas mediante el método de la regresión, se les aplico la prueba de correlación de Spearman que se presenta en la Tabla 4; en la cual se puede observar que no existe relación entre las puntuaciones para los rendimientos de los 6 cursos a un nivel de significación del 1%, lo queindica que no existen problemas de multicolinealidad.

Tabla 4. Correlaciones de los factores.

			FAC_RM	FAC_RV	FAC_MAT	FAC_FIS	FAC_QUI	FAC_BIO
Rho de Spearman	FAC_RM	Coeficiente de correlación	1	-0.018	0.011	-0.01	0	-0.01
		Sig. (bilateral)	.	0.273	0.515	0.518	0.991	0.542
		N	3799	3799	3799	3799	3799	3799
	FAC_RV	Coeficiente de correlación	-0.018	1	0.005	-0.011	-0.01	0.017
		Sig. (bilateral)	0.273	.	0.76	0.516	0.531	0.304
		N	3799	3799	3799	3799	3799	3799
	FAC_MAT	Coeficiente de correlación	0.011	0.005	1	-0.026	0.004	0.006
		Sig. (bilateral)	0.515	0.76	.	0.111	0.829	0.72
		N	3799	3799	3799	3799	3799	3799
	FAC_FIS	Coeficiente de correlación	-0.01	-0.011	-0.026	1	-0.005	-0.008
		Sig. (bilateral)	0.518	0.516	0.111	.	0.743	0.639
		N	3799	3799	3799	3799	3799	3799
	FAC_QUI	Coeficiente de correlación	0	-0.01	0.004	-0.005	1	-0.006
		Sig. (bilateral)	0.991	0.531	0.829	0.743	.	0.724
		N	3799	3799	3799	3799	3799	3799
	FAC_BIO	Coeficiente de correlación	-0.01	0.017	0.006	-0.008	-0.006	1
		Sig. (bilateral)	0.542	0.304	0.72	0.639	0.724	.
		N	3799	3799	3799	3799	3799	3799

Fuente: Elaboración Propia

Análisis de las variables explicativas

Pedret [17], recomienda hacer un análisis previo de las variables explicativas, antes de estimar la función discriminante. La Tabla 5 muestra la prueba de igualdad de medias de los grupos de ingresantes y no ingresantes en cada variable independiente.

Entre las variables que discriminan adecuadamente a un nivel de significación del 1% se encuentran los rendimientos de los cursos de Razonamiento Matemático, Razonamiento Verbal, Matemática y Biología. Además la primera variable a ingresar en el modelo sería el rendimiento en Razonamiento Matemático ya que presenta el valor estadístico F más alto (1683.764) y el lambda de Wilks (0.693) más bajo, de esta manera se justificó la presencia indispensable esta variable.

Tabla 5. Prueba de Igualdad de Medias de los grupos.

	Lambda de Wilks	F	gl1	gl2	Sig.
FAC_RM	0.693	1683.764	1	3797	0
FAC_RV	0.998	8.159	1	3797	0.004
FAC_MAT	0.992	31.633	1	3797	0
FAC_FIS	0.999	3.795	1	3797	0.051
FAC_QUI	1	0.826	1	3797	0.363
FAC_BIO	0.991	33.165	1	3797	0

Fuente: Elaboración Propia

Función discriminante lineal de Fisher

Para la obtención de la función discriminante lineal de Fisher se realizó la diferencia entre las funciones de Ingreso-No Ingreso delaTabla 6.

$$Z_i = -1.3269 + 2.3020FAC\_RM - 0.1923FAC\_RV - 0.3775FAC\_MAT + 0.1312FAC\_FIS - 0.0613FAC\_QUI + 0.3865FAC\_BIO$$

Decisión:

Si $Z_i < 0$, se clasifica al individuo "i" en el grupo formado por los no ingresantes.

Si $Z_i > 0$, se clasifica al individuo "i" en el grupo formado por los ingresantes.

Tabla 6. Coeficientes de la Función Discriminante Lineal de Fisher.

	Ingreso	
	No Ingreso	Ingreso
FAC_RM	-0.347	1.955
FAC_RV	0.029	-0.163
FAC_MAT	0.057	-0.321
FAC_FIS	-0.02	0.111
FAC_QUI	0.009	-0.052
FAC_BIO	-0.058	0.328
(Constante)	-0.736	-2.063

Fuente: Elaboración Propia

En la Tabla 7 se muestra el estadístico Chi-cuadrado correspondiente al Lambda de Wilks para contrastar si la función discriminante es significativa, se reportó un P-valor=0.000 lo que indicó que posee un buen poder de clasificación para los ingresantes y no ingresantes.

Tabla 7. Lambda de Wilks.

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	0.673	1505.252	6	0

Fuente: Elaboración Propia

En la Tabla 8 se aprecia la matriz de estructura, que indicó que elcurso que posee una mayor capacidad discriminante es Razonamiento Matemático, ya que tiene una alta correlación con la función discriminante; seguido por Biología y Matemática.

Tabla 8. Matriz de estructura.

	Función 1
FAC_RM	0.954
FAC_BIO	0.134
FAC_MAT	-0.131
FAC_RV	-0.066
FAC_FIS	0.045
FAC_QUI	-0.021

Fuente: Elaboración Propia

La función discriminante Lineal de Fisher clasificó correctamente al 83% postulantes. En la Tabla 9 se puede observar que la clasificación correcta para los no ingresantes fue del 80.8%, mientras que para los ingresantes fue del 95.5%.

Tabla 9. Resultados de clasificación.

		Ingreso	Grupo de pertenencia pronosticado		Total
			No Ingreso	Ingreso	
Original	Recuento	No Ingreso	2606	620	3226
		Ingreso	26	547	573
	%	No Ingreso	80.8	19.2	100
		Ingreso	4.5	95.5	100

Fuente: Elaboración Propia

Validación Cruzada

Posteriormente se aplicó la validación cruzada en 10 grupos. Los resultados proporcionados en la Tabla 10indicaron que la función discriminante lineal de Fisher predijo satisfactoriamente al 82.7% de postulantes; donde la predicción correcta para los ingresantes fue del 95.2% y para los no ingresantes del 80.4%.

Tabla 10. Predicción mediante Validación Cruzada.

Muestra	Condición		
	No Ingresante	Ingresante	Total
1	82.70%	94.70%	84.50%
2	82.90%	94.80%	84.70%
3	79.40%	100.00%	82.40%
4	85.60%	97.50%	86.80%
5	81.30%	96.20%	83.40%
6	79.20%	95.60%	82.10%
7	80.30%	94.00%	82.10%
8	83.90%	88.90%	84.70%
9	79.70%	93.80%	82.10%
10	69.40%	96.90%	74.10%
Validación Cruzada	80.40%	95.20%	82.70%

Fuente: Elaboración Propia

Análisis discriminante con Algoritmos Genéticos
Generación de la Población Inicial

Se aplicó un remuestreo a las puntuaciones factoriales, generándose 20 muestras aleatorias a las que se les aplicó el Análisis Discriminante. Fue así como se obtuvo 20 funciones discriminantes lineales de Fisher generadas por el software estadístico SPSS.

Método de la ruleta

En la Tabla 11 se presentan las 5 funciones discriminantes seleccionadas (1, 4, 6, 17 y 20) con errores promedio de clasificación que oscilaron entre 0.1145 y 0.1221.

Tabla 11. Funciones discriminantes seleccionadas.

N°	RM	RV	MAT	FIS	QUI	BIO	(Constante)	Error romedio
1	2.3181	-0.2606	-0.3664	0.0508	-0.1282	0.396	-1.3803	0.1221
4	2.401	-0.0854	-0.3495	0.0826	-0.0696	0.3957	-1.5131	0.1145
6	2.2668	-0.1439	-0.3546	0.1706	-0.0383	0.3043	-1.2863	0.1154
17	2.3733	-0.2612	-0.3849	0.0102	-0.0162	0.4298	-1.4128	0.1206
20	2.3384	-0.2688	-0.3228	0.1928	0.0165	0.2221	-1.2968	0.118

Fuente: Elaboración Propia

Funciones obtenidas mediante cruce y mutación

La aplicación de los operadores genéticos se realizó a través de funciones elaboradas en el software R.

Se utilizó las cinco funciones de la Tabla 11 para realizar el cruce aritmético, tomando en cuenta que se necesitan dos funciones progenitoras para generar dos funciones nuevas o hijas, en este caso se generaron 20 ($2C_2^5 = 20$). Se observó que el error promedio de clasificación tuvo valores entre 0.1160 y 0.1221, cumpliendo con la tolerancia solo aquellas funciones con errores menores o iguales al error promedio de la función discriminante lineal de Fisher (0.1185), las cuales se seleccionaron y se muestran en la Tabla 12.

Tabla 12. Funciones obtenidas del cruce que cumplen con la tolerancia.

N°	RM	RV	MAT	FIS	QUI	BIO	(Constante)	Error 1	Error 2	Error Promedio
4	2.3582	-0.1664	-0.3573	0.072	-0.0946	0.392	-1.4441	0.1761	0.0558	0.116
6	2.3276	-0.1536	-0.3559	0.1083	-0.0728	0.3605	-1.3909	0.1813	0.0506	0.116
12	2.286	-0.1701	-0.3573	0.1366	-0.0622	0.3315	-1.3206	0.1888	0.0436	0.1162
11	2.3356	-0.1968	-0.3602	0.0741	-0.1008	0.385	-1.407	0.1801	0.0524	0.1162
5	2.3217	-0.1549	-0.356	0.113	-0.0707	0.3559	-1.3809	0.1823	0.0506	0.1164
2	2.3652	-0.1611	-0.3568	0.0689	-0.0949	0.3958	-1.4557	0.1745	0.0593	0.1169
20	2.3143	-0.2084	-0.3676	0.0904	-0.0488	0.3666	-1.3494	0.1888	0.0454	0.1171
16	2.2779	-0.1691	-0.3572	0.1447	-0.0577	0.3241	-1.3066	0.1906	0.0436	0.1171
7	2.3666	-0.2053	-0.3685	0.0446	-0.0646	0.4095	-1.4355	0.1779	0.0576	0.1178
8	2.3672	-0.1996	-0.3673	0.0471	-0.0668	0.4082	-1.4386	0.1779	0.0576	0.1178
1	2.3539	-0.1849	-0.3591	0.0645	-0.1029	0.3959	-1.4377	0.1767	0.0593	0.118

Fuente: Elaboración Propia

Los coeficientes que presentaron mayor variabilidad fueron los correspondientes a los cursos de Física y Química (38.76% y 24.8%). En consecuencia, se decidió no tomar como las mejores soluciones las funciones anteriores, que presentaron un menor error a comparación de función discriminante lineal de Fisher, y se continuó con el algoritmo con la finalidad de encontrar funciones con error aún más pequeño.

Se utilizó el conjunto de funciones dadas por el cruce aritmético para realizar la mutación uniforme. Para ello se seleccionó de forma aleatoria una columna, en este caso fue la séptima (constante de la función), de la cual se obtuvo su valor mínimo y máximo, los que fueron considerados para obtener un valor aleatorio con distribución uniforme, el cual resultó -1.4215.

Como se alteró el error promedio, se conservó solo las funciones que cumplieron con la tolerancia, obteniendo como resultado 8 funciones discriminantes que se muestran en la Tabla 13.

Tabla 13. Funciones obtenidas de la mutación que cumplen con la tolerancia.

N°	RM	RV	MAT	FIS	QUI	BIO	(Constante)	Error 1	Error 2	Error Promedio
4	2.3582	-0.1664	-0.3573	0.072	-0.0946	0.392	-1.4215	0.1789	0.0524	0.1156
6	2.3276	-0.1536	-0.3559	0.1083	-0.0728	0.3605	-1.4215	0.1764	0.0541	0.1152
12	2.286	-0.1701	-0.3573	0.1366	-0.0622	0.3315	-1.4215	0.1754	0.0558	0.1156
5	2.3217	-0.1549	-0.356	0.113	-0.0707	0.3559	-1.4215	0.1767	0.0541	0.1154
2	2.3652	-0.1611	-0.3568	0.0689	-0.0949	0.3958	-1.4215	0.1807	0.0524	0.1165
20	2.3143	-0.2084	-0.3676	0.0904	-0.0488	0.3666	-1.4215	0.1773	0.0541	0.1157
16	2.2779	-0.1691	-0.3572	0.1447	-0.0577	0.3241	-1.4215	0.1748	0.0541	0.1145
1	2.3539	-0.1849	-0.3591	0.0645	-0.1029	0.3959	-1.4215	0.1789	0.0558	0.1174

Fuente: Elaboración Propia

Función discriminante óptima

Se continuó con el algoritmo genético hasta llegar a la séptima generación, ya que en esta etapa el error promedio se estabilizó.

En cada generación el error promedio fue disminuyendo como se puede observar en la Figura3.

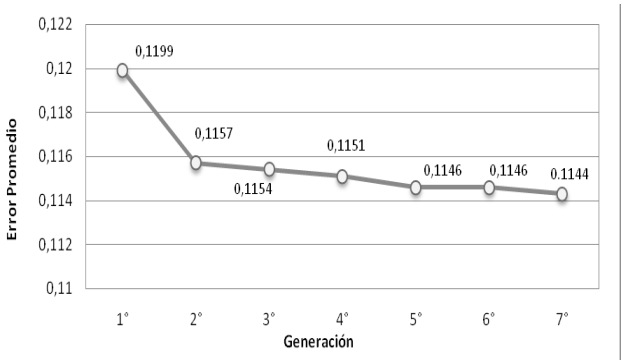

Figura 3. Evolución del Análisis Discriminante con Algoritmos Genéticos.
Fuente: Elaboración Propia

En la séptima generación se obtuvo 6 funciones discriminantes que proporcionaron el mismo error mínimo de clasificación.

Debido a que se encontró un conjunto de soluciones factibles, se tomó solo una de ellas, siendo la función discriminante óptima:

La función discriminante lineal con Algoritmos Genéticos clasificó correctamente al 84.2% postulantes. La clasificación correcta para los no ingresantes fue del 82.4%, mientras que para los ingresantes fue del 94.8%.

Validación Cruzada en Algoritmos Genéticos

Se aplicó la Validación Cruzada en 10 grupos. Los resultados que se presentan en la Tabla 14indicaron que la función discriminante lineal con Algoritmos Genéticos predijo satisfactoriamente al 83.1% de postulantes, donde la predicción correcta para los ingresantes fue del 95.3% y para los no ingresantes del 80.9%.

Tabla 14. Predicción mediante Validación Cruzada en Algoritmos Genéticos.

Muestra	Condición		
	No Ingresante	Ingresante	Total
1	83.30%	94.70%	85.00%
2	81.40%	94.80%	83.40%
3	80.10%	100.00%	82.90%
4	84.40%	95.00%	85.50%
5	82.30%	96.20%	84.20%
6	80.80%	97.10%	83.70%
7	82.10%	94.00%	83.70%
8	84.90%	90.50%	85.80%
9	79.70%	93.80%	82.10%
10	70.10%	96.90%	74.70%

Continuación de tabla 14

Validación Cruzada	80.90%	95.30%	83.10%

Fuente: Elaboración Propia

Comparación de resultados

La comparación de ambos métodos se realizó mediante los porcentajes de error de clasificación y predicción. Para la predicción, se utilizó la validación cruzada en 10 grupos cuyo procedimiento se explicó en la sección de revisión de literatura.

Se presenta la Tabla 15 como resumen de la comparación de ambas técnicas.

Tabla 15. Comparación de porcentajes de error de clasificación y predicción.

	Análisis Discriminante Lineal de Fisher			Análisis Discriminante con Algoritmos Genéticos		
Condición	No Ingresantes	Ingresantes	Total	No Ingresantes	Ingresantes	Total
Clasificación	19.20%	4.50%	17.00%	17.60%	5.20%	15.80%
Predicción	19.60%	4.80%	17.30%	19.10%	4.70%	16.90%

Fuente: Elaboración Propia

Se puede observar que el porcentaje de error de clasificación mejora al utilizar la técnica de Análisis discriminante con Algoritmos Genéticos que el Análisis discriminante lineal de Fisher, disminuyendo de 17.0% a un 15.8%. De la misma forma para la predicción, disminuyendo de 17.3% a 16.9%. Esta mejora alrededor del 1% significó un aumento en la clasificación y predicción correcta de aproximadamente 38 postulantes.

4. Conclusiones

El Análisis discriminante con algoritmos genéticos encontró una función discriminante que no sólo permite clasificar mejor a un postulante sino que también predice la condición del mismo, brindando una tasa de error de clasificación y predicción de 15.8% y 16.9% respectivamente, los cuales son menores a los obtenidos con el Análisis discriminante lineal de Fisher, que brinda un 17.0% y 17.3%.

El Análisis discriminante lineal determinó que el puntaje correspondiente al curso de Razonamiento Matemático posee una mayor capacidad discriminante seguido por Biología y Matemática (Álgebra, Aritmética, Geometría y Trigonometría) en el perfil de rendimiento de los postulantes; lo que guarda relación con sus pesos en el examen de admisión, ya que juntos representan más del 50% de preguntas de toda la evaluación, y además representan más del 60% de cursos dictados en el centro de estudios preuniversitarios.

5. Literatura citada

Back, B; Laitinen, T; Sere, K y Wezel, M. 1996. Chooosing bankruptcy predictors using discriminant analysis, logit analysis, and genetic algorithms. p. 4.

Efron, B y Tibshirani, R.1993. An Introduction to the bootstrap CHAPMAN. p. 239-240.

Gestal, M. 2010. Introducción a los Algoritmos Genéticos. Universidad de Coruña. p. 9-14.

Gil, N. 2006. Algoritmos Genéticos. Universidad de Colombia, Escuela de Estadística sede Medellín. p. 22.

Goldberg, D. 1989. Genetic Algorithms in Search, Optimización and Machine Learning. Addison-Wesley. 412 p.

Hair, Anderson;Tatham; Black.2008 Análisis Multivariante. 5 ed. PEARSON. p. 249-279.

Hastie, T; Tibshirani, R; Friedman, J. 2008. The elements of statistical learning.2 ed. SPRINGER. p. 241-245.

Hernández, O. 1998. Temas de análisisestadísticomultivariado.1 ed. Universidad de Costa rica. p. 136-138

Holland, J. 1975. Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor, Michigan. 183 p.

Johnson, D. 2004. Métodos multivariados aplicados al análisis de datos. THOMSON. p. 217-274.

Koza, J. 1992. Genetic Programming. On the Programming of Computers by Means of Natural Selection. MIT Press. 819 p.

Manly, B. 1986. Multivariate Statistical Methods. CHAPMAN. p. 122-123.

Manrique, D. Computación Evolutiva: Algoritmos geneticos. Universidad Politécnica de Madrid. p. 11-12.

Montanero, J. 2008. Análisis multivariante. Universidad de Extremadura. p. 230-231

Montano, A; Cantú, M. 2011. Algoritmos Genéticos en la discriminación. 136 p.

Moujahid, A; Inza, I; Larrañaga, P. 2008. Algoritmos Genéticos. Universidad del País Vasco. p. 1.

Pedret, R; Sagnier, L; Camp, F. 2000. Herramientas para segmentar mercados y posicionar productos. 2 ed. DEUSTO. p. 228-234.

Reeves, C. 2010. Genetic Algorithms. School of Mathematical and Information Sciences. p. 63-64.

Rosas, F. 2000. Reconocimiento de patrones de rendimiento de los postulantes en el concurso de admisión 2005-I de la Universidad Nacional Agraria La Molina usando la técnica Análisis discriminante.42 p.

Sharma, S. 1996. AppliedMultivariateTechniques. WILEY. p. 263-264.

Tolmos, P. 2003. Introducción a los algoritmos genéticos y sus aplicaciones. Universidad Rey Juan Carlos, Servicio de Publicaciones. p. 6.

Uriel, Ezequiel; Aldás, J. 2005. Análisis Multivariante Aplicado. 1 ed. THOMSON. p. 278-309.