UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

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

Presentado por:

JOAO MANUEL RADO HUARINGA

TESIS PARA OPTAR EL TÍTULO DE

INGENIERO ESTADÍSTICO E INFORMÁTICO

Lima-Perú

2014

ÍNDICE GENERAL

I. INTRODUCCIÓN 1

II. REVISIÓN DE LITERATURA 2

2.1 El Análisis Discriminante Lineal 2

2.1.1 Definición del análisis discriminante lineal 2

2.1.2 Objetivos del análisis discriminante lineal 3

2.1.3 Función Discriminante Lineal de Fisher 3

2.1.4 Estimación de la función discriminante 4

2.1.5 Supuestos del Análisis Discriminante Lineal 6

2.2 Algoritmos Genéticos 7

2.2.1 Definición de Algoritmos Genéticos 7

2.2.2 Objetivos de los Algoritmos Genéticos 8

2.2.3 Cromosoma 8

2.2.4 Población Inicial 8

2.2.5 Función de aptitud 9

2.2.6 Selección 9

2.2.7 Operadores genéticos 9

2.2.7.1 Reproducción 10

2.2.7.2 Cruce 10

2.2.7.3 Mutación 10

2.3 Análisis discriminante con algoritmos genéticos 10

2.3.1 Generación de la población inicial 10

2.3.2 Cálculo de la función de aptitud 11

2.3.3 Selección por el método de la ruleta 11

2.3.4 Cruce aritmético 12

2.3.5 Mutación uniforme 12

2.4 Metodología del Algoritmo Genético con Análisis Discriminante 13

2.5 Validación cruzada en K grupos 14

2.6 Indicador de comparación de técnicas 15

2.6.1 Tasa de error de clasificación 15

2.7 El Centro de Estudios Preuniversitarios (CEPRE-UNALM) 15

2.7.1 ¿Qué es el CEPRE-UNALM? 15

2.7.2 Régimen académico 16

2.7.2.1 Estrategias 16
2.7.2.2 Cursos 16
2.7.2.3 Antecedentes del perfil de rendimiento 17
III. MATERIALES Y MÉTODOS 18
3.1 Materiales y equipo 18
3.2 Metodología de la investigación 18
3.2.1 Tipo de la investigación 18
3.2.2 Diseño de la investigación 18
3.2.3 Instrumento de colecta de datos 18
3.2.4 Formulación de hipótesis 19
3.2.5 Identificación de variables 19
3.2.6 Definiciones operacionales 20
3.2.7 Población y Muestra 20
3.3 Metodología aplicada 21
IV. RESULTADOS Y DISCUSIÓN 22
4.1 Análisis estadístico univariado 22
4.2 Análisis estadístico bivariado 23
4.3 Análisis discriminante lineal 25
4.3.1 Verificación de supuestos 25
4.3.2 Análisis de las variables explicativas 27
4.3.3 Función discriminante lineal de Fisher 28
4.3.4 Validación cruzada 30
4.4 Análisis discriminante con algoritmos genéticos 31
4.4.1 Generación de la población inicial 31
4.4.2 Método de la ruleta 31
4.4.3 Funciones obtenidas mediante cruce y mutación 32
4.4.4 Función discriminante óptima 34
4.4.5 Validación cruzada en Algoritmos genéticos 35
4.5 Comparación de resultados 36
V. CONCLUSIONES 37
VI. RECOMENDACIONES 38
VII. REFERENCIAS BIBLIOGRÁFICAS 39
VIII. ANEXOS 41
Anexo I: Funciones discriminantes de la primera iteración 41

Anexo II: Programas en R para la aplicación de algoritmos genéticos 44
Anexo III: Aplicación del Análisis Discriminante Lineal 47
Anexo IV: Programas en R para la aplicación de algoritmos genéticos 48

ÍNDICE DE CUADROS

Cuadro N° 1. Estadísticos de grupo 22
Cuadro N° 2. Prueba de Homogeneidad de Varianzas de Levene 26
Cuadro N° 3. Prueba M de box 26
Cuadro N° 4. Correlaciones de los factores 27
Cuadro N° 5. Prueba de Igualdad de Medias de los grupos 28
Cuadro N° 6. Coeficientes de la Función Discriminante Lineal de Fisher 28
Cuadro N° 7. Lambda de Wilks 29
Cuadro N° 8. Matriz de estructura 29
Cuadro N° 9. Resultados de clasificación 30
Cuadro N° 10. Predicción mediante Validación Cruzada 30
Cuadro N° 13. Funciones discriminantes seleccionadas 31
Cuadro N° 15. Funciones obtenidas del cruce que cumplen con la tolerancia 32
Cuadro N° 17. Funciones obtenidas de la mutación que cumplen con la tolerancia 33
Cuadro N° 18. Funciones discriminantes óptimas 34
Cuadro N° 19. Predicción mediante Validación Cruzada en Algoritmos Genéticos 35
Cuadro N° 20. Comparación de porcentajes de error de clasificación y predicción 36
Cuadro N° 21. Datos para la aplicación del Análisis Discriminante 47
Cuadro N° 22. Funciones discriminantes generadas y su error de clasificación 48
Cuadro N° 23. Funciones discriminantes y su tasa de error promedio de clasificación 49
Cuadro N° 24. Funciones discriminantes y su ajuste 50
Cuadro N° 25. Selección de individuos del Análisis discriminante con Algoritmos Genéticos 50
Cuadro N° 26. Funciones discriminantes para la aplicación del cruce aritmético 51
Cuadro N° 27. Funciones discriminantes obtenidas del cruce aritmético entre las funciones N° 2 y 3 52
Cuadro N° 28. Funciones discriminantes obtenidas del Cruce aritmético 52
Cuadro N° 29. Funciones discriminantes obtenidas con la Mutación uniforme 53
Cuadro N° 30. Funciones discriminantes óptimas 54

ÍNDICE DE GRÁFICOS

Gráfico N° 1. Estructura de la evolución de un Algoritmo Genético 13
Gráfico N° 2. Gráfico de Dispersión para Ingresantes 23
Gráfico N° 3. Gráfico de Dispersión para No Ingresantes 24
Gráfico N° 4. Evolución del Análisis Discriminante con Algoritmos Genéticos 34
Gráfico N° 5. Generación de la población inicial 48
Gráfico N° 6. Cruce Aritmético 51

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 claves: 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 Nacional Agraria 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.

Keywords: Discriminant Analysis, Genetic Algorithms, Optimization.

I. INTRODUCCIÓN

En ciencia Estadística se han venido desarrollando técnicas de análisis multivariado con 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 se han 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 universidad y cuya preparación se realizó en su Centro de Estudios Preuniversitarios (CEPRE_UNALM).

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

II. REVISIÓN DE LITERATURA

2.1 El Análisis Discriminante Lineal

2.1.1 Definición del análisis discriminante lineal

Pedret [17] indica que “el análisis discriminante permite determinar cuáles son las variables (de entre la serie de variables seleccionadas previamente por el investigador), que mejor explican la pertenencia de un individuo a un determinado grupo. Además logra determinar el grupo al que pertenecerá un individuo pendiente de clasificación, basándose en la respuesta o valores que toma dicho individuo en la serie de variables que más explican la pertenencia a cada grupo”.

Según Uriel [22], “el análisis discriminante se utiliza para clasificar a distintos individuos en grupos, o poblaciones, alternativos de los valores de un conjunto de variables sobre los individuos que se pretende clasificar; donde cada individuo pertenece a un solo grupo. La pertenencia a un grupo u otro se introduce mediante una variable categórica (dependiente) que toma tantos valores como grupos existentes”.

Johnson [10], hace una comparación entre el análisis de regresión y el análisis discriminante: “El análisis discriminante es semejante al de regresión, excepto que la variable dependiente es categórica, en lugar de continua. En el análisis de regresión se desea poder predecir el valor de una variable de interés con base en un conjunto de variables predictoras. En el análisis discriminante se desea poder predecir la pertenencia a una clase de una observación particular, con base en un conjunto de variables predictoras.”

Sharma [20], hace hincapié que el análisis discriminante lineal de Fisher requiere la verificación de los supuestos de normalidad multivariada y homocedasticidad. Además, indica que de no cumplirse estos supuestos, los resultados de clasificación se verían afectados.

Finalmente Manly [12] explica que “el Análisis discriminante es bastante robusto a la violación de los supuestos mencionados anteriormente. Sin embargo, al interpretar los resultados el investigador debe ser consciente de los posibles efectos debido a la violación de los supuestos”.

2.1.2 Objetivos del análisis discriminante Lineal

Según Hair, Anderson, Tatham y Black [6], "El análisis discriminante puede tratar cualquiera de los siguientes objetivos de investigación:

- Determinar si existen diferencias estadísticamente significativas entre los perfiles de las puntuaciones medias sobre un conjunto de variables de dos (o más) grupos definidos a priori.
- Determinar cuál de las variables independientes cuantifica mejor las diferencias en los perfiles de las puntuaciones medias de dos o más grupos.
- Establecer los procedimientos para clasificar objetos (individuos, empresas, productos, etc.), dentro de los grupos, en base a sus puntuaciones sobre un conjunto de variables independientes.
- Establecer el número y la composición de las dimensiones de la discriminación entre los grupos formados a partir del conjunto de variables independientes".

Por otro lado Johnson [10] señala que: "El objetivo básico del análisis discriminante es producir una regla o un esquema de clasificación que permita a un investigador predecir la población de la que es lo más probable que tenga que venir una observación."

2.1.3 Función Discriminante Lineal de Fisher

Pedret [17], define la función discriminante lineal de la siguiente manera: "Si la variable a explicar es de m grupos, el análisis discriminante calcula m-1 funciones discriminantes. La estimación de la función discriminante se efectúa reduciendo las variables explicativas iniciales a unas nuevas variables, combinaciones lineales de las primeras. Los valores tomados por estas nuevas variables se llaman puntuaciones discriminantes. Cada individuo obtiene una puntuación discriminante en cada una de las funciones discriminantes".

Si llamamos Z_i a la puntuación discriminante asociada al individuo i $(i = 1, \dots, n)$ en una función discriminante cualquiera, Z_i será una combinación lineal de las variables explicativas iniciales X_p $(p = 1, \dots, P)$:

$$Z_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \dots + b_p X_{pi} \qquad \forall i = 1, \dots, n$$

Siendo b_p el coeficiente discriminante o peso asociado a la variable X_p.

2.1.4 Estimación de la función discriminante

Uriel [22], explica el sustento matemático para la obtención de la función discriminante de Fisher, como una función lineal de K variables explicativas X:

$$D = u_1 X_1 + u_2 X_2 + ... + u_K X_K \tag{2.1}$$

El problema planteado es la obtención de los coeficientes de ponderación u_j. Si se considera la existencia de n observaciones, la función discriminante se expresa de la siguiente manera:

$$D_i = u_1 X_{1i} + u_2 X_{2i} + ... + u_K X_{Ki} \tag{2.2}$$

Así, D_i es la puntuación discriminante correspondiente a la observación i-ésima.

La expresión en forma matricial es:

$$\begin{bmatrix} D_1 \\ D_2 \\ ... \\ D_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} & ... & X_{K1} \\ X_{12} & X_{22} & ... & X_{K2} \\ ... & ... & ... & ... \\ X_{1n} & X_{2n} & ... & X_{Kn} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ ... \\ u_K \end{bmatrix} \rightarrow d = Xu \tag{2.3}$$

La variabilidad de la función discriminante se puede expresar de la siguiente manera:

$$d'd = u'X'Xu \tag{2.4}$$

El segundo miembro de (2.4) es una forma cuadrática de la matriz X'X. Esta matriz, al estar expresadas las variables en desviaciones respecto a la media, es la matriz de suma de cuadrados y productos cruzados (SCPC) total de las variables X. Esta matriz se puede descomponer como una matriz SCPC entre grupos y SCPC intra-grupos:

$$X'X = T = F + W \tag{2.5}$$

donde T, F y W son las matrices de SCPC total, entre-grupos e intra-grupos respectivamente. Sustituyendo (2.5) en (2.4), se obtiene:

$$d'd = u'Tu = u'Fu + u'Wu \tag{2.6}$$

Para la estimación de los coeficientes u_j, Fisher utilizó el siguiente criterio:

$$\textit{Máximización de } \lambda = \frac{u'Fu}{u'Wu} \tag{2.7}$$

Se trata de que el primer término (Variabilidad entre-grupos) sea lo mayor posible en detrimento del segundo término (Variabilidad intra-grupos).

Los centros de gravedad o centroides (vector de medias), son los estadísticos básicos que resumen la información sobre los grupos. Sustituyendo en (2.2) $X_1, X_2, ..., X_K$ por los elementos del vector de medias, se obtiene:

$$D_I = u_1 \bar{X}_{1,I} + u_2 \bar{X}_{2,I} + ... + u_K \bar{X}_{K,I} \tag{2.8}$$

Produciendo de forma análoga en el grupo II:

$$D_{II} = u_1 \bar{X}_{1,II} + u_2 \bar{X}_{2,II} + ... + u_K \bar{X}_{K,II} \tag{2.9}$$

El punto de corte discriminante C se calcula promediando, es decir:

$$C = \frac{\bar{D}_I + \bar{D}_{II}}{2} \tag{2.10}$$

El criterio para clasificar al individuo i es el siguiente:

Si $D_i < C$, se clasifica al individuo i en el grupo I

Si $D_i > C$, se clasifica al individuo i en el grupo II

En general, cuando se aplica el análisis discriminante se le resta el valor de C a la función. De esta forma, la función discriminante viene dada por:

$$D - C = u_1 X_1 + u_2 X_2 + ... + u_K X_K - C \tag{2.11}$$

Existe una forma alternativa a la utilización (2.11), que consiste en construir funciones discriminantes para cada grupo, basadas también en el criterio (2.7). Estas funciones, a las que denominaremos, tienen la siguiente estructura:

$$\begin{aligned} F_I &= a_{I,1} X_1 + a_{I,2} X_2 + ... + a_{I,K} X_K - C_I \\ F_{II} &= a_{II,1} X_1 + a_{II,2} X_2 + ... + a_{II,K} X_K - C_{II} \end{aligned} \tag{2.12}$$

Cuando se utilizan estas funciones, se clasifica a un individuo en el grupo para el que la función F_j sea mayor. A partir de los coeficientes de las funciones (2.12) se pueden obtener los coeficientes de la función (2.11), a través de la siguiente equivalencia:

$$\begin{aligned} F_{II} - F_I &= (a_{II,1} - a_{I,1})X_1 + (a_{II,1} - a_{I,1})X_2 + ... + (a_{II,K} - a_{I,K})X_K - (C_{II} - C_I) \\ &= u_1 X_1 + u_2 X_2 + ... + u_K X_K - C = D - C \end{aligned} \tag{2.13}$$

2.1.5 Supuestos del Análisis Discriminante Lineal

Según Hair, Anderson, Tatham y Black [6], para obtener la función discriminante se requiere del supuesto de normalidad multivariante de las variables independientes. Los datos que no cumplan con este supuesto pueden causar problemas en la estimación de la función discriminante.

Otro supuesto que se requiere es la homogeneidad de matrices varianza covarianza de los grupos. La violación de este supuesto puede afectar a la clasificación, causando problemas de “sobreclasificación” dentro de los grupos de matrices varianza covarianza grandes.

Un problema a tomar en cuenta es la multicolinealidad entre las variables independientes. Esto ocurre cuando dos o más variables están altamente correlacionadas provocando poca capacidad explicativa al conjunto completo.

Por último, un supuesto implícito es que todas las relaciones son lineales. Por lo que las no lineales no se encuentran reflejadas en la función discriminante.

Montanero [14], explica que “el método lineal de clasificación de Fisher manifiesta ser bastante robusto frente a violaciones moderadas de estas hipótesis. Por ello en el caso (frecuente) de que no se verifiquen las mismas, no debemos de descartar la estrategia lineal sino que hemos de ejecutarla y evaluar su validez a posteriori. Si los resultados no son satisfactorios, optaremos por otro tipo de clasificación, como la cuadrática”. Además, indica que una forma de evaluar la validez es a través de los métodos de Jackniffe o validación cruzada.

2.2 Algoritmos Genéticos

2.2.1 Definición de Algoritmos Genéticos

Goldberg [5] señala que los algoritmos genéticos son “algoritmos de búsqueda basados en los mecanismos de selección natural y genética natural. Combinan la supervivencia de los más compatibles entre las estructuras de cadenas, con una estructura de información ya aleatorizada, intercambiada para construir un algoritmo de búsqueda con algunas de las capacidades de innovación de la búsqueda humana”.

Moujahid, Inza y Larrañaga [16], detallan que “los algoritmos genéticos son métodos que ayudan a resolver problemas de búsqueda y optimización. Están basados en el proceso genético de los organismos vivos. A través de generaciones, los individuos evolucionan en la naturaleza de acorde a los principios de selección natural y la supervivencia de los más fuertes, según Darwin (1859). Como aprendizaje de este proceso, los algoritmos genéticos trabajan con una población de individuos, cada uno de los cuales representa una solución factible a un problema planteado. Estos individuos serán seleccionados, evaluados y reproducidos tras el paso de cada generación. Propagándose de esta manera características a lo largo de cada una de ellas y evolucionando hasta obtener la solución requerida”.

Finalmente Koza [11] brinda una definición más formal de los algoritmos genéticos: “El algoritmo genético es un algoritmo matemático que transforma un conjunto de objetos matemáticos individuales con respecto al tiempo, usando operaciones modeladas de acuerdo al principio de Darwin de reproducción y supervivencia del más apto y tras haberse presentado de forma natural una serie de operaciones genéticas de entre las que destaca la recombinación sexual. Cada uno de estos objetos matemáticos suele ser una cadena de caracteres (letras o números) de longitud fija que se ajusta al modelo de las cadenas de cromosomas, y se les asocia con una cierta función matemática que refleja su ajuste (fitness)”.

2.2.2 Objetivos de los Algoritmos Genéticos

Holland [9], brinda un punto de vista computacional señalando que los objetivos de los algoritmos genéticos son:

- Imitar los procesos adaptativos de los sistemas naturales
- Diseñar sistemas artificiales (programas) que retengan los mecanismos importantes de los sistemas naturales.

Por su parte Gil [4], hace incidencia en la optimización: “Los algoritmos genéticos son, simplificando, algoritmos de optimización, es decir; tratan de encontrar la mejor solución a un problema dado entre un conjunto de soluciones posibles”.

Manrique [13], señala el objetivo como un sistema robusto de la computación evolutiva: “Los problemas del mundo real casi nunca son estáticos y los problemas de optimización temporal son cada vez más comunes. Estas circunstancias requieren un cambio en la estrategia que se aplica para resolver el problema. La potencia de los algoritmos genéticos viene de que la técnica que usan es robusta y puede tratar con éxito un gran número de tipos de problemas, incluyendo y destacando aquellos que son difícilmente solucionables utilizando otro tipo de métodos clásicos”.

Finalmente Montano [15] señala que “Los algoritmos genéticos se basan en esquemas formando grupos de individuos; donde se localizan los óptimos, para posteriormente encontrar la función óptima de las variables aleatorias”.

2.2.3 Cromosoma

Según Back [1], “un cromosoma o individuo es una solución candidata al problema que se desea resolver, los cromosomas son llamados también cadenas, las que a su vez se componen de un número de genes, los cuales llevan valores llamados alelos”.

2.2.4 Población Inicial

Es un conjunto inicial de cromosomas. Los aspectos que se debe tener en cuenta en la población son: tamaño y generación.

Respecto al tamaño que debe tomar, Reeves [18], señala que “una población pequeña no permitirá explorar el espacio de búsqueda de manera efectiva, pero si es demasiado

grande hay posibilidad de que la eficiencia del método disminuya y en consecuencia no encontrar una solución óptima en un plazo razonable de tiempo."

Por otro lado, Gil [4], basándose en una evidencia empírica, sugiere que un tamaño de población comprendida entre el I y 2I es suficiente para atacar con éxito el problema planteado. Siendo I la longitud de una ristra (conjunto de genes).

Finalmente Montano [15], indica que la población inicial debe estar formada entre 20 y 100 cromosomas.

La generación de la población se explica en la sección 2.3.1

2.2.5 Función de Aptitud

Según información recopilada de internet, la función de aptitud es la función objetivo del problema de optimización. Se caracteriza por ser capaz de "castigar" a las malas soluciones, y de "premiar" a las buenas, de forma que sean estas últimas las que se propaguen con mayor rapidez.

Tolmos [21], explica la función de aptitud en términos de capacidad frente al problema a solucionar: "El algoritmo suele requerir una función de capacidad o potencial que asigna una puntuación (capacidad) a cada cromosoma de la población actual. La capacidad o el potencial de un cromosoma depende de cómo resuelva ese cromosoma el problema a tratar".

Finalmente Gestal [3], brinda una definición más formal explicando que "es una medida numérica de la bondad de una solución. Indica si los individuos de la población representan o no buenas soluciones al problema planteado".

2.2.6 Selección

Según Gestal [3], "Los algoritmos de selección serán los encargados de escoger qué individuos van a disponer de oportunidades de reproducirse y cuáles no". De esta forma aumenta la posibilidad de tener buenos individuos en un futuro.

2.2.7 Operadores Genéticos

Según Montano [15], los operadores genéticos proporcionan los mecanismos de búsqueda básicos de los algoritmos genéticos; los cuales se usan para crear nuevas soluciones,

basadas en el conjunto de las mejores soluciones escogidas previamente por algún método de selección.

2.2.7.1 Reproducción

Es un proceso en el cual el cromosoma (cadena) es copiado a la nueva generación. El cromosoma con un alto valor de ajuste tiene mayor oportunidad de participar en la próxima generación".

2.2.7.2 Cruce

El cruce toma dos individuos y produce dos nuevos individuos. Una parte de un cromosoma es combinada con otra parte de otro cromosoma. En esta operación existe la posibilidad de combinar las partes buenas de dos cromosomas, proporcionando la descendencia de dos nuevos cromosomas, al menos iguales a los progenitores".

2.2.7.3 Mutación

El objetivo de este operador es introducir nuevo material genético en la población, o mínimo prevenir la pérdida de éste. Bajo la mutación, un gen puede recibir un valor que no ocurre antes en la población, o que tiene que perderse debido a la reproducción".

Gestal [3], hace hincapié en la utilización de este operador junto al de cruce, "la mutación de un individuo provoca que uno de sus genes, varíe su valor de forma aleatoria, produciendo un nuevo individuo. Además se suele utilizar de manera conjunta con el operador de cruce".

2.3 Análisis discriminante con algoritmos genéticos

El Análisis Discriminante Lineal con algoritmos genéticos, es la combinación del análisis discriminante lineal y los algoritmos genéticos. Según Montano [15], tiene como objetivo principal clasificar individuos con el mínimo error posible. Se caracteriza por utilizar las variables discriminantes resultantes del análisis discriminante lineal y operadores genéticos, lo cual permite estimar una función o un conjunto de funciones discriminantes.

2.3.1 Generación de Población Inicial

Según Gil [4], "la población inicial de un Algoritmo Genético puede ser creada de muy diversas formas, desde generar aleatoriamente el valor de cada gen para cada individuo,

utilizar una función ávida o generar alguna parte de cada individuo y luego aplicar una búsqueda local".

Montano [15], explica una metodología para la generación de la población inicial: "Se generan 2 muestras…Se toma una de ellas como primera muestra y se le aplica la técnica de remuestreo…, con la finalidad de obtener n nuevas muestras del mismo tamaño; mientras que la segunda se reserva, porque su información se usará para evaluar"

"…Posteriormente a cada una de las muestras obtenidas por remuestreo se le aplicó análisis discriminante de manera que se obtienen n funciones discriminantes (cromosomas)".

2.3.2 Cálculo de la función de Aptitud

Montano [15], señala que para calcular el fitness o función de aptitud se deben realizar los siguientes pasos:

1. Para cada una de las funciones se obtiene la proporción de individuos mal clasificados en cada grupo, posteriormente se realiza la suma, la cual será dividida entre 2, obteniéndose el error promedio de clasificación.
2. Ya que se tiene el error promedio de cada función se realiza la suma, generándose el error promedio total del problema.
3. La aptitud es igual al error promedio de clasificación entre el error promedio total del problema.

2.3.3 Selección por el método de la ruleta

Según Montano [15], la selección se realiza aleatoriamente, y se asigna una probabilidad P_j a cada individuo j, perteneciente a la población en cuestión y se toma como base el valor de ajuste. También se generan números aleatorios con distribución uniforme, uno para cada elemento de la población, y se comparan contra la probabilidad acumulada $c_i = \sum_{j=1}^{i} P_j$. Entonces, la solución o cromosoma i se selecciona para integrar la nueva población si $c_{i-1} < U(0,1) < c_i$.

En el método de la ruleta, las funciones progenitoras son seleccionadas de acuerdo a la probabilidad asignada (P_j), tomando como referencia su valor de ajuste. Se tiene que el

individuo con mayor proporción será seleccionado más veces que el de menor proporción. La probabilidad P_i para cada individuo es definida por:

$$P_i = \frac{F_i}{\sum_{i=1}^{N} F_i} \tag{2.14}$$

donde F_i, es el valor de ajuste del i-ésimo individuo y N es el tamaño de la población. Este método limita al algoritmo genético a encontrar la optimización de la solución debido a la asignación de probabilidades.

2.3.4 Cruce Aritmético

Sean S_m , y S_k , dos soluciones o cromosomas, entonces al seleccionar aleatoriamente un valor r de una variable aleatoria con distribución U(0, 1) , se logra la generación de dos nuevas soluciones o nuevos cromosomas al considerar las combinaciones convexas entre los dos, es decir:

$$S'_m = rS_m + (1-r)S_k \quad \text{y} \quad S'_k = rS_k + (1-r)S_m \tag{2.15}$$

De esta forma se generan dos nuevas soluciones.

2.3.5 Mutación Uniforme

Sea S(i), una matriz de tamaño s x p cuyas filas representan posibles soluciones del problema en la iteración i con p componentes, y

$$a_j = \min_i S_{i,j} \quad \text{y} \quad b_j = \max_i S_{i,j} \quad \text{donde } j = 1,..., p$$

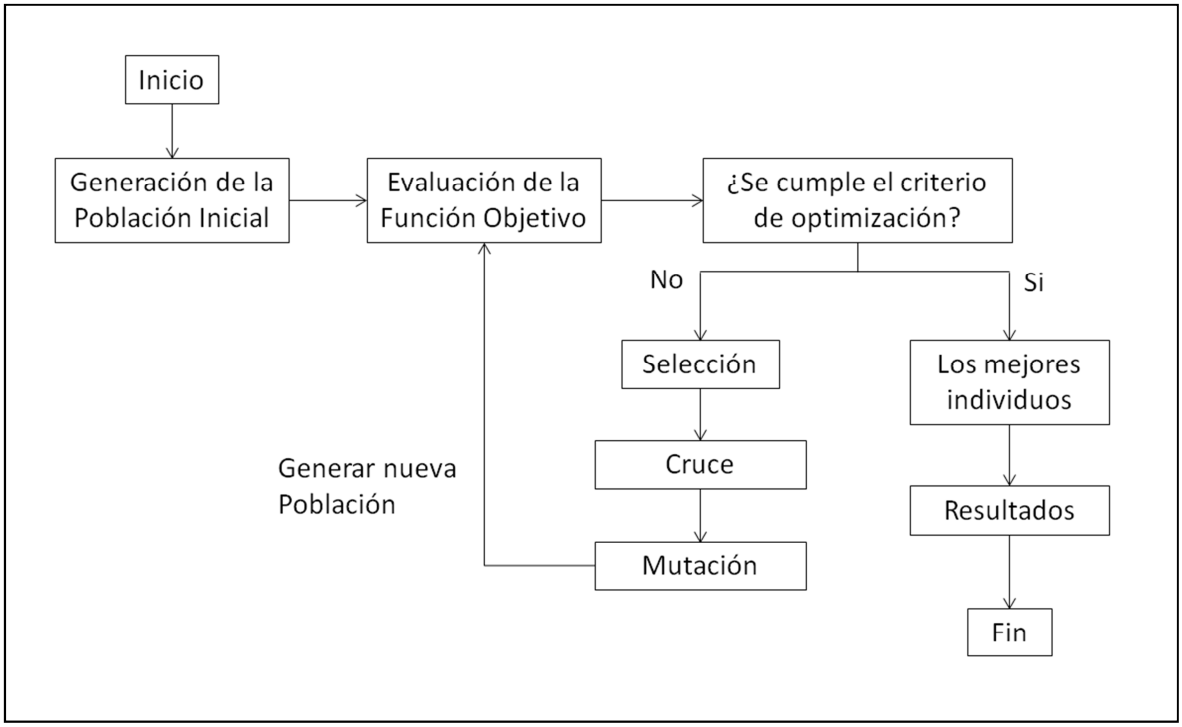
Es decir, el límite inferior y superior del conjunto de cada columna de la matriz S(i) del problema. Seleccionando aleatoriamente el valor k entre 1 y p y un valor de la variable aleatoria $U(a_k, b_k)$ con distribución uniforme en el intervalo (a_k, b_k), se realiza la siguiente mutación:

$$\begin{aligned} S_{i,j} &= U(a_k, b_k) \quad i = 1,..., p \quad \text{si } j = k \\ S_{i,j} &= S_{i,j} \qquad\quad i = 1,..., p \quad \text{si } j \neq k \end{aligned} \tag{2.16}$$

2.4 Metodología del Algoritmo Genético con Análisis Discriminante

Para la aplicación de los algoritmos genéticos al Análisis Discriminante Lineal, se utiliza los pasos mostrados en el Gráfico N° 1.

Gráfico N° 1. Estructura de la evolución de un Algoritmo Genético.



Fuente: Algoritmos genéticos en la discriminación – Montano (2011)

Montano [15], indica que se debe comenzar generando la población inicial con un conjunto de funciones discriminantes. Posteriormente se evalúa cada una de ellas, para realizar esto se sustituye los vectores de las X's de la muestra que generó la población inicial, luego se contabilizan los casos mal clasificados para los grupos en cada función, lo cual se utiliza para determinar la proporción de error de clasificación y se conservan las funciones que reportan menor error (error cero).

El siguiente paso es asignar la función de aptitud (probabilidad de mejor ajuste) a cada una de las funciones discriminantes. Seguidamente se procede a la selección de individuos, aplicando el método de la ruleta, es decir, se generan n números aleatorios con distribución uniforme ordenados de manera ascendente y se obtiene el ajuste acumulado, si el valor se encuentra entre los valores acumulados, entonces el individuo es seleccionado. Se continúa hasta el n-ésimo valor aleatorio.

Posteriormente se realiza el cruce aritmético y la mutación uniforme. Después de aplicar cada operador se procede a evaluar las funciones con las variables independientes de la muestra que generó la población inicial.

Se toma como tolerancia el error de clasificación obtenido aplicando sólo el análisis discriminante lineal de Fisher (1936), esperando lograr obtener una o un conjunto de funciones que reporten un menor error de clasificación que al aplicar sólo el análisis discriminante.

Si se encuentran funciones óptimas en el proceso de cruce o mutación, estas se conservan para validarlas con la muestra de prueba. En el caso de que no se hayan logrado obtener funciones óptimas el algoritmo se repite volviendo a la selección de individuos.

2.5 Validación Cruzada en k grupos

Efron y Tibshirani [2], explican el algoritmo de la siguiente manera:

1. Dividir los datos en K partes aproximadamente del mismo tamaño.
2. En la K-ésima parte, ajustar el modelo con el resto de datos (K-1) y calcular el error de predicción del modelo ajustado, tomando como datos de prueba el conjunto de datos K.
3. Hacer lo anterior para K=1,2,..,k y combinar las K estimaciones de error de predicción.

El error de la validación cruzada se obtiene mediante:

$$CV = \frac{1}{k}\sum_{k=1}^{K} E_k \tag{2.17}$$

Hastie, Tibshirani y Friedman [7] muestran mediante una curva de aprendizaje, obtenida de comparar la tasa de clasificación correcta frente al tamaño de muestra de entrenamiento, que para valores de K igual a 5 o 10 se obtiene una menor varianza y una mejora en la clasificación a mayor tamaño de muestra. Además que no existen cambios sustanciales en esta tasa si la muestra es mayor a 100. Finalmente concluyen que: “...Si la curva de aprendizaje tiene una pendiente considerable frente al tamaño de los datos de entrenamiento, utilizar un K igual a 5 o 10 sobreestima el error de predicción. Presentando como inconveniente un sesgo que se deja a criterio del investigador. En comparación con la validación cruzada dejando uno fuera, se presenta un sesgo bajo pero una mayor variabilidad. En general, se recomienda tomar valores de K igual a 5 o 10: ver Brieman y Spector (1992) y Kohavi (1995)”.

2.6 Indicador de comparación de técnicas

2.6.1 La Tasa de error de clasificación

Según Hernández [8], “un procedimiento para medir la eficacia de cualquier regla de clasificación consiste en calcular su “tasa de error”, o probabilidad total de clasificación errónea. Esta puede calcularse directamente cuando las poblaciones son conocidas completamente (forma y parámetros). Sin embargo, esto no ocurre con frecuencia en la práctica; lo usual es que algunos de los parámetros deban estimarse basándose en una muestra de cada población”.

La tasa de error de clasificación se obtiene de aplicar la función discriminante estimada a todos los casos de la muestra. Luego, se cuentan los casos donde la clasificación fue incorrecta, basándose en la información del grupo de pertenencia de cada uno.

La tasa de error de clasificación o la tasa de error aparente, es la división del número de casos mal clasificados por la función discriminante estimada entre el total de casos. Una expresión para esta tasa es la siguiente:

$$\text{Tasa de error de clasificación} = \frac{\text{N° de casos mal clasificados}}{\text{N° total de casos}} \tag{2.18}$$

Montanero [14], indica que las funciones discriminantes con tasas de error de clasificación menores al 25% son consideradas aceptables. Se debe tomar en cuenta que las tasas de error de clasificación son útiles cuando se disponen de grandes tamaños de muestra en cada población.

2.7 El Centro de Estudios Preuniversitarios (CEPRE-UNALM)

2.7.1 ¿Qué es el CEPRE-UNALM?

Según el Artículo 438° del Reglamento General de la UNALM, el Centro de Estudios Pre Universitarios UNALM es una dependencia de la Universidad Nacional Agraria La Molina y tiene la finalidad de brindar a sus alumnos la formación necesaria para ingresar y seguir con éxito su carrera profesional en la universidad. Para su funcionamiento recibe el apoyo del cuerpo docente de la universidad. Depende funcionalmente del vicerrectorado académico.

Según el Artículo 2° del Reglamento de Organización y Funciones del Centro de Estudios Preuniversitarios UNALM, el CEPRE-UNALM tiene los siguientes objetivos:

a) Impartir una seria y rigurosa preparación para ingresar a la universidad a través de las modalidades de ingreso directo y concurso general de admisión.
b) Contribuir significativamente al aprendizaje de los conocimientos necesarios para el éxito en los estudios universitarios.

2.7.2 Régimen académico

2.7.2.1 Estrategias

Según información de su página web, algunas de las estrategias que ha ido aplicando el CEPRE-UNALM para contribuir al logro de sus objetivos son:

a) Desarrollo de una sólida organización pedagógica y administrativa.
b) Enfoque en aspectos formativos y reforzamiento de técnicas de estudio.
c) Realización de asesorías y orientaciones permanentes.
d) Orientación a una educación personalizada.
e) Preparación a través de grupos de estudio y seminarios conducidos por profesores especializados.
f) Preferencia en el ingreso a su cuerpo docente de profesores nombrados o contratados de la UNALM.
g) Actualización de sus guías y materiales de estudio.

2.7.2.2 Cursos

En el CEPRE-UNALM se dictan 9 cursos: Razonamiento Matemático, Razonamiento Verbal, Algebra, Aritmética, Geometría, Trigonometría, Biología, Química y Física. Estos cursos son los que se evalúan en el examen de admisión de la universidad. Y además, cumplen con el syllabus detallado en el prospecto.

2.7.2.3 Antecedentes del perfil de rendimiento

- **Perfil de postulantes al examen de admisión 2015-I de la UNALM cuya preparación se realizó en el CEPRE-UNALM**

Rosas [19], determina la existencia de un perfil diferenciado relativamente mayor en el rendimiento de los postulantes al concurso de admisión 2005-I, que se prepararon en el CEPRE-UNALM respecto de los que no se prepararon en ella. Mediante una aplicación del análisis discriminante en cuatro grupos (Pertenece al CEPRE y Si Ingresó, Pertenece al CEPRE y No Ingresó, No pertenece al CEPRE y Si Ingresó, No pertenece al CEPRE y No Ingresó), concluyó que fue el curso de Química el que tuvo mayor capacidad discriminante y a la vez la mayor participación de docentes con vinculación laboral con la UNALM en la plana docente del Centro de estudios preuniversitarios. Además, que en los grupos de postulantes que si ingresaron los alumnos del CEPRE-UNALM se encontraron mejor clasificados (66.7%) que los no se prepararon en ella (60.8%), en el caso de los no ingresantes el comportamiento fue de forma inversa.

III. MATERIALES Y MÉTODOS

3.1 Materiales y Equipo

- Una computadora Toshiba Intel Corei3 de 64 bits.
- Una Impresora hp Laser Jet P1102w.
- Una tinta color negro.
- Un millar de hojas bond A4.
- Software Minitab 16, SPSS versión 19 y R versión 3.0.3.

3.2 Metodología de la Investigación

3.2.1 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.

3.2.2 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.

3.2.3 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.

3.2.4 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.

3.2.5 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 dependiente y tiene dos categorías:

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

Las variables independientes o predictoras lo 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.

3.2.6 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.

3.2.7 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.

3.3 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.

IV. RESULTADOS Y DISCUSIÓN

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

1. Análisis estadístico univariado

Cuadro N° 1. Estadísticos de grupo

INGRESO		Media	Desv. típ.	N válido (según lista)	
				No ponderados	Ponderados
No Ingreso	RM	8,8143	3,50320	3226	3226,000
	RV	8,9110	3,37889	3226	3226,000
	MAT	6,1921	4,16514	3226	3226,000
	FIS	4,9235	3,89772	3226	3226,000
	QUI	8,3511	5,07840	3226	3226,000
	BIO	4,9860	3,95024	3226	3226,000
Ingreso	RM	12,9012	2,97240	573	573,000
	RV	11,4629	3,17956	573	573,000
	MAT	12,8660	2,99597	573	573,000
	FIS	10,8508	3,25120	573	573,000
	QUI	14,3593	3,08542	573	573,000
	BIO	9,4624	3,89469	573	573,000
Total	RM	9,4308	3,72710	3799	3799,000
	RV	9,2959	3,47150	3799	3799,000
	MAT	7,1987	4,66790	3799	3799,000
	FIS	5,8175	4,35811	3799	3799,000
	QUI	9,2573	5,28751	3799	3799,000
	BIO	5,6612	4,25463	3799	3799,000

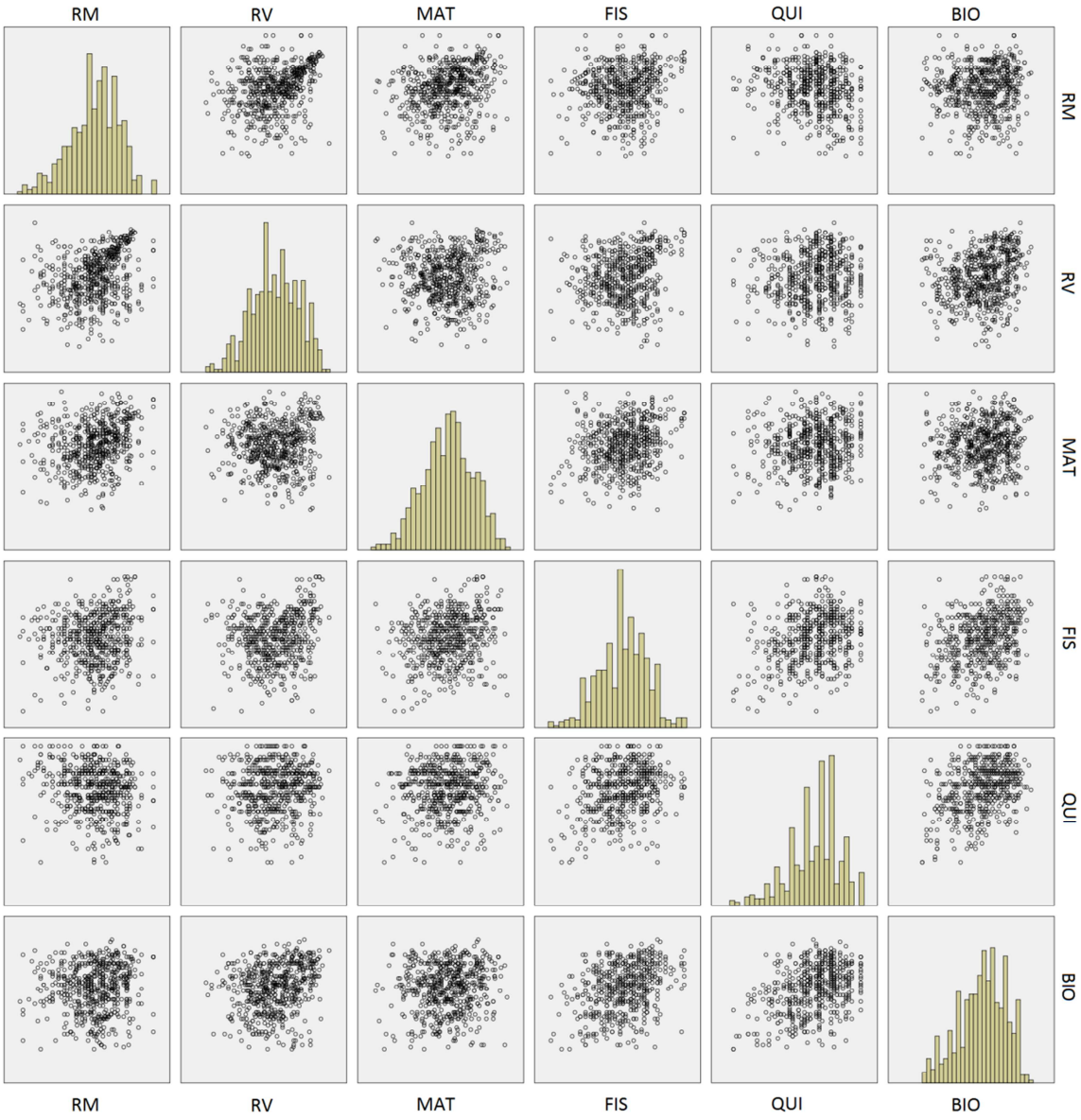
Fuente: Elaboración Propia

En el Cuadro N° 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.

2. Análisis estadístico bivariado

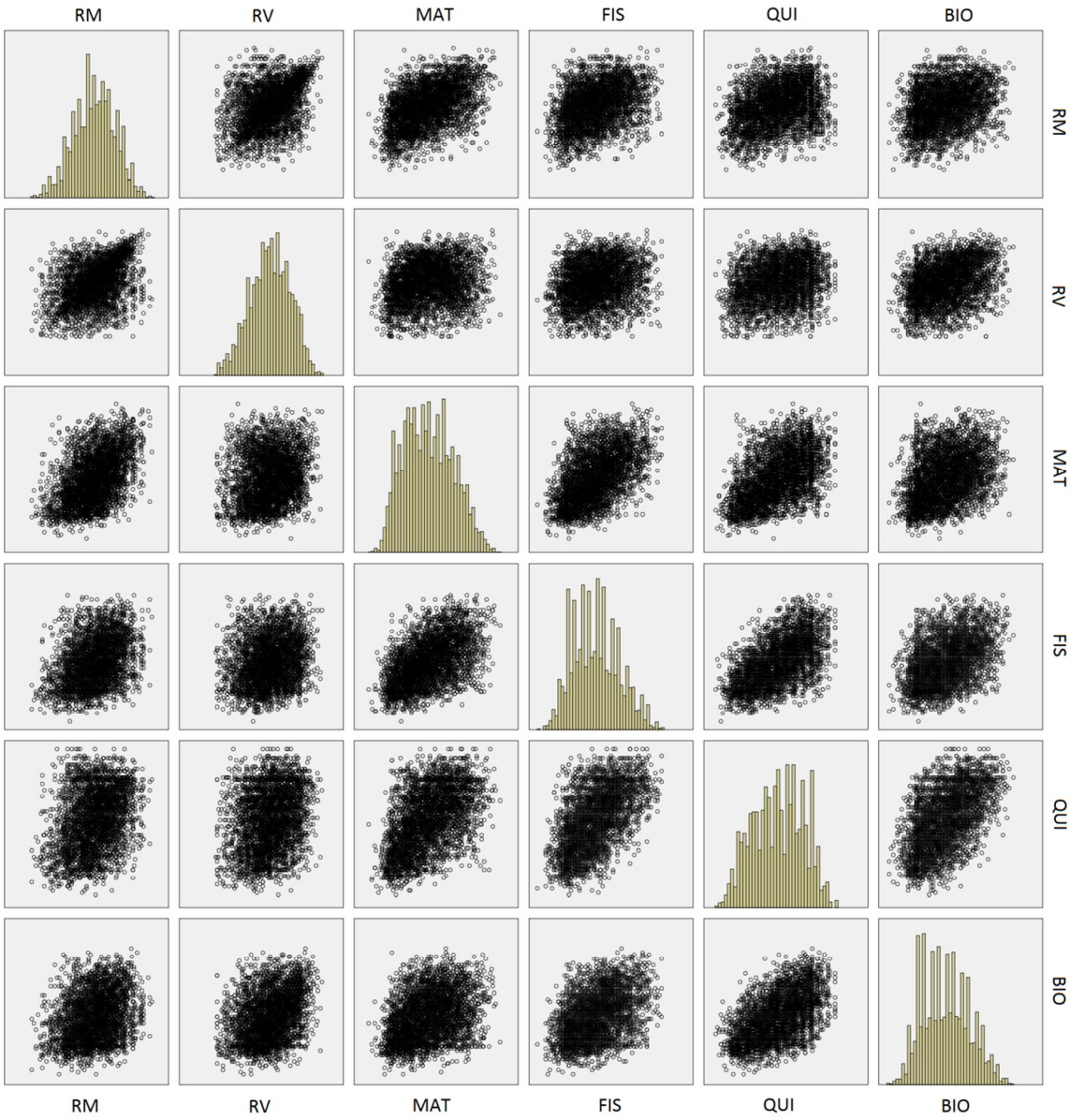
En los Gráficos N° 2 y N° 3 se puede observar que los rendimientos se encuentran correlacionados en los 6 cursos para los ingresantes y no ingresantes. Este resultado da un indicio de que se encuentrae problemas de multicolinealidad.

Gráfico N° 2. Gráfico de Dispersión para Ingresantes



Fuente: Elaboración Propia

Gráfico N° 3. Gráfico de Dispersión para No Ingresantes



Fuente: Elaboración Propia

3. Análisis discriminante lineal

3.1 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 los Gráficos N° 2 y N° 3 se 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 Shapiro Wilk 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 el Cuadro N° 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.

Cuadro N° 2. Prueba de Homogeneidad de Varianzas de Levene

Prueba de homogeneidad de varianzas

	Estadístico de Levene	gl1	gl2	Sig.
RM	29,039	1	3797	,000
RV	2,159	1	3797	,142
MAT	111,344	1	3797	,000
FIS	38,918	1	3797	,000
QUI	232,096	1	3797	,000
BIO	1,465	1	3797	,226

Fuente: Elaboración Propia

En el Cuadro N° 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.

Cuadro N° 3. Prueba M de box

Resultados de la prueba

M de Box		347,186
F	Aprox.	16,470
	gl1	21
	gl2	3778544,944
	Sig.	,000

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 el Cuadro N° 4; en el 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 que indica que no existen problemas de multicolinealidad.

Cuadro N° 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	1,000	-,018	,011	-,010	,000	-,010
		Sig. (bilateral)	.	,273	,515	,518	,991	,542
		N	3799	3799	3799	3799	3799	3799
	FAC_RV	Coeficiente de correlación	-,018	1,000	,005	-,011	-,010	,017
		Sig. (bilateral)	,273	.	,760	,516	,531	,304
		N	3799	3799	3799	3799	3799	3799
	FAC_MAT	Coeficiente de correlación	,011	,005	1,000	-,026	,004	,006
		Sig. (bilateral)	,515	,760	.	,111	,829	,720
		N	3799	3799	3799	3799	3799	3799
	FAC_FIS	Coeficiente de correlación	-,010	-,011	-,026	1,000	-,005	-,008
		Sig. (bilateral)	,518	,516	,111	.	,743	,639
		N	3799	3799	3799	3799	3799	3799
	FAC_QUI	Coeficiente de correlación	,000	-,010	,004	-,005	1,000	-,006
		Sig. (bilateral)	,991	,531	,829	,743	.	,724
		N	3799	3799	3799	3799	3799	3799
	FAC_BIO	Coeficiente de correlación	-,010	,017	,006	-,008	-,006	1,000
		Sig. (bilateral)	,542	,304	,720	,639	,724	.
		N	3799	3799	3799	3799	3799	3799

Fuente: Elaboración Propia

3.2 Análisis de las variables explicativas

Pedret [17], recomienda hacer un análisis previo de las variables explicativas, antes de estimar la función discriminante. El Cuadro N° 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.

Cuadro N° 5. Prueba de Igualdad de Medias de los grupos

	Lambda de Wilks	F	gl1	gl2	Sig.
FAC_RM	,693	1683,764	1	3797	,000
FAC_RV	,998	8,159	1	3797	,004
FAC_MAT	,992	31,633	1	3797	,000
FAC_FIS	,999	3,795	1	3797	,051
FAC_QUI	1,000	,826	1	3797	,363
FAC_BIO	,991	33,165	1	3797	,000

Fuente: Elaboración Propia

3.3 Función discriminante lineal de Fisher

Para la obtención de la función discriminante lineal de Fisher (3.19) se realizó la diferencia entre las funciones de Ingreso-No Ingreso del Cuadro N° 6.

$$Z_i = -1.3269 + 2.3020FAC\_RM - 0.1923FAC\_RV - 0.3775FAC\_MAT + 0.1312FAC\_FIS - 0.0613FAC\_QUI + 0.3865FAC\_BIO \quad (2.19)$$

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.

Cuadro N° 6. Coeficientes de la Función Discriminante Lineal de Fisher

	INGRESO	
	No Ingreso	Ingreso
FAC_RM	-,347	1,955
FAC_RV	,029	-,163
FAC_MAT	,057	-,321
FAC_FIS	-,020	,111
FAC_QUI	,009	-,052
FAC_BIO	-,058	,328
(Constante)	-,736	-2,063

Fuente: Elaboración Propia

En el Cuadro N° 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.

Cuadro N° 7. Lambda de Wilks

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	,673	1505,252	6	,000

Fuente: Elaboración Propia

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

Cuadro N° 8. Matriz de estructura

	Función
	1
FAC_RM	,954
FAC_BIO	,134
FAC_MAT	-,131
FAC_RV	-,066
FAC_FIS	,045
FAC_QUI	-,021

Fuente: Elaboración Propia

La función discriminante Lineal de Fisher clasificó correctamente al 83% postulantes. En el Cuadro N° 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%.

Cuadro N° 9. Resultados de clasificación

			Grupo de pertenencia pronosticado		
		INGRESO	No Ingreso	Ingreso	Total
Original	Recuento	No Ingreso	2606	620	3226
		Ingreso	26	547	573
	%	No Ingreso	80,8	19,2	100,0
		Ingreso	4,5	95,5	100,0

Fuente: Elaboración Propia

3.4 Validación Cruzada

Posteriormente se aplicó la validación cruzada en 10 grupos. Los resultados proporcionados en el Cuadro N° 10 indicaron 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%.

Cuadro N° 10. Predicción mediante Validación Cruzada

Muestra	**Condición**		
	No Ingresante	**Ingresante**	**Total**
1	82.7%	94.7%	84.5%
2	82.9%	94.8%	84.7%
3	79.4%	100.0%	82.4%
4	85.6%	97.5%	86.8%
5	81.3%	96.2%	83.4%
6	79.2%	95.6%	82.1%
7	80.3%	94.0%	82.1%
8	83.9%	88.9%	84.7%
9	79.70%	93.80%	82.10%
10	69.40%	96.90%	74.10%
Validación Cruzada	**80.4%**	**95.2%**	**82.7%**

Fuente: Elaboración Propia

4. Análisis discriminante con Algoritmos Genéticos

4.1 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. La población inicial se puede observar en el Cuadro N° 11 (Ver Anexo I.), las primeras 8 columnas corresponden al número de función y a los coeficientes de las funciones discriminantes. Las siguientes columnas muestran los errores de reclasificar el conjunto de datos con cada función discriminante; la columna 9, el error de clasificación general, las 10 y 11, el error de clasificación del grupo 1 (No ingresantes) y grupo 2 (Ingresantes), y la última columna el error promedio de clasificación de ambos grupos.

4.2 Método de la ruleta

En el Cuadro N° 12 (Ver Anexo I.) se presenta los resultados obtenidos al aplicar la selección mediante el método de la ruleta, la columna 3 muestra la probabilidad de ajuste, seguida por la probabilidad de ajuste acumulada, la selección de la función y los números aleatorios con distribución uniforme ordenados de menor a mayor.

La selección consistió en ver si el valor aleatorio generado se encontraba entre las probabilidades de ajuste acumuladas (PA_{i-1} y PA_i) de la función i. En el Cuadro N° 13 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.

Cuadro N° 13. Funciones discriminantes seleccionadas

n	RM	RV	MAT	FIS	QUI	BIO	(Constante)	Error Promedio
1	2.3181	-0.2606	-0.3664	0.0508	-0.1282	0.3960	-1.3803	0.1221
4	2.4010	-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	0.1180

Fuente: Elaboración Propia

4.3 Funciones obtenidas mediante cruce y mutación

La aplicación de los operadores genéticos se realizó a través de funciones elaboradas en el software R. (Ver Anexo II)

Se utilizó las cinco funciones del Cuadro N° 13 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 ($2C_2^5 = 20$), las que se presentan en el Cuadro N° 14 (Ver Anexo I.)

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 el Cuadro N° 15.

Cuadro N°15. 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	0.0720	-0.0946	0.3920	-1.4441	0.1761	0.0558	0.1160
6	2.3276	-0.1536	-0.3559	0.1083	-0.0728	0.3605	-1.3909	0.1813	0.0506	0.1160
12	2.2860	-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	0.3850	-1.4070	0.1801	0.0524	0.1162
5	2.3217	-0.1549	-0.3560	0.1130	-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	0.1180

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.

Las funciones resultantes de la mutación se pueden observar en el Cuadro N° 16 (Ver Anexo I.) 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 el Cuadro N° 17.

Cuadro N°17. 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	0.0720	-0.0946	0.3920	-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	2.2860	-0.1701	-0.3573	0.1366	-0.0622	0.3315	-1.4215	0.1754	0.0558	0.1156
5	2.3217	-0.1549	-0.3560	0.1130	-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

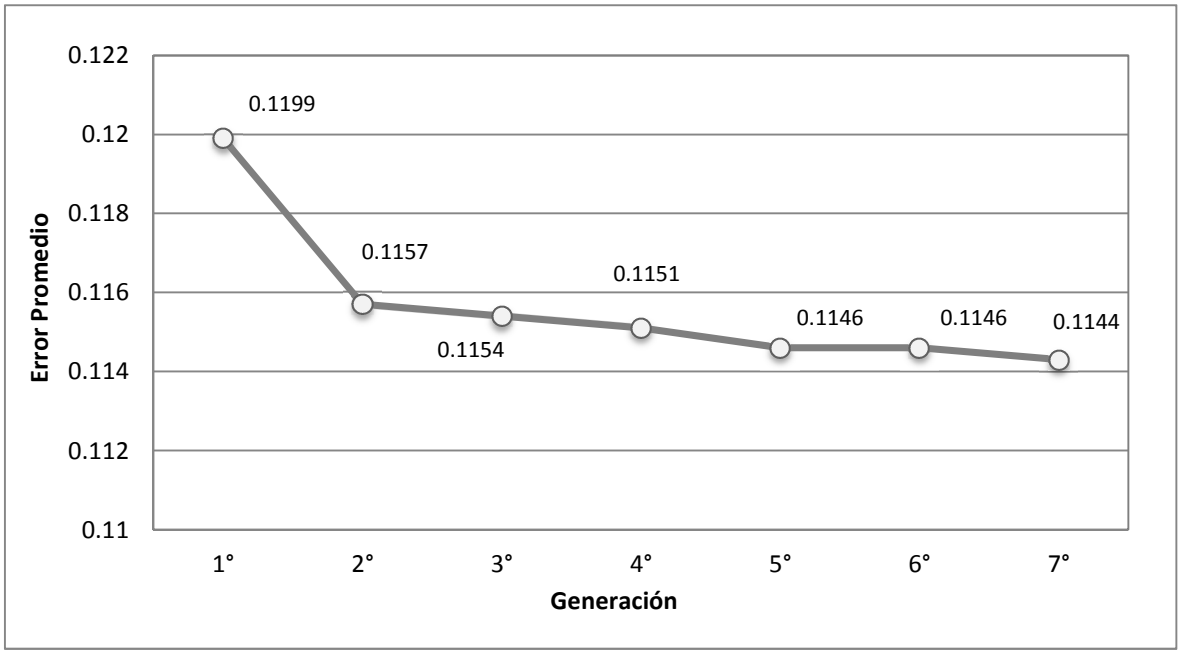
Fuente: Elaboración Propia

4.4 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 el Gráfico N° 4.

Gráfico N° 4. 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. Las cuales son presentadas en el Cuadro N°18.

Cuadro N° 18. Funciones discriminantes óptimas

N°	RM	RV	MAT	FIS	QUI	BIO	(Constante)	Error 1	Error 2	Error Promedio
1	2.3194	-0.1858	-0.3634	0.0992	-0.0614	0.3634	-1.4215	0.1764	0.0524	0.1144
2	2.3194	-0.1858	-0.3634	0.0992	-0.0614	0.3649	-1.4215	0.1764	0.0524	0.1144
3	2.3194	-0.1858	-0.3634	0.0992	-0.0614	0.3644	-1.4215	0.1764	0.0524	0.1144
4	2.3194	-0.1858	-0.3634	0.0992	-0.0614	0.3626	-1.4215	0.1764	0.0524	0.1144
5	2.3194	-0.1858	-0.3634	0.0992	-0.0614	0.3632	-1.4215	0.1764	0.0524	0.1144
6	2.3194	-0.1858	-0.3634	0.0992	-0.0614	0.3620	-1.4215	0.1764	0.0524	0.1144

Fuente: Elaboración Propia

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

$$Z_i = -1.4215 + 2.3194 FAC\_RM - 0.1858 FAC\_RV - 0.3634 FAC\_MAT + 0.0992 FAC\_FIS - 0.0614 FAC\_QUI + 0.3634 FAC\_BIO \quad (2.20)$$

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%.

4.5 Validación Cruzada en Algoritmos Genéticos

Se aplicó la Validación Cruzada en 10 grupos. Los resultados que se presentan en el Cuadro N° 19 indicaron 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%.

Cuadro N° 19. Predicción mediante Validación Cruzada en Algoritmos Genéticos

Muestra	Condición		
	No Ingresante	Ingresante	Total
1	83.3%	94.7%	85.0%
2	81.4%	94.8%	83.4%
3	80.1%	100.0%	82.9%
4	84.4%	95.0%	85.5%
5	82.3%	96.2%	84.2%
6	80.8%	97.1%	83.7%
7	82.1%	94.0%	83.7%
8	84.9%	90.5%	85.8%
9	79.7%	93.8%	82.1%
10	70.1%	96.9%	74.7%
Validación Cruzada	**80.9%**	**95.3%**	**83.1%**

Fuente: Elaboración Propia

5. 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 el Cuadro N° 20 como resumen de la comparación de ambas técnicas.

Cuadro N° 20. 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	19.2%	4.5%	**17.0%**	17.6%	5.2%	**15.8%**
Predicción	19.6%	4.8%	**17.3%**	19.1%	4.7%	**16.9%**

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.

V. CONCLUSIONES

1. 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%.

2. 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.

3. El porcentaje de error de clasificación en ingresantes obtenida por el Análisis discriminante con algoritmos genéticos (5.2%) reportó un aumento en comparación con el Análisis discriminante lineal de Fisher (4.5%). Esto se dio a causa del desigual tamaño que tuvieron los grupos de ingresantes y no ingresantes.

4. En la aplicación de la mutación uniforme en el algoritmo genético, se encontró que al tomar a uno de los coeficientes de los cursos que poseen capacidad discriminante, se dieron cambios sustanciales en el error promedio de clasificación.

5. En la aplicación del análisis discriminante no se cumplieron los supuestos para estimar la función discriminante. Sin embargo, según la literatura consultada se prosiguió con la técnica para aplicar validación cruzada, donde no se encontraron problemas en la predicción de postulantes.

6. Por otro lado, la función discriminante obtenida mediante los algoritmos genéticos puede clasificar a nuevos postulantes, empleando el puntaje obtenido en los seis cursos del examen de admisión. De esta forma se puede discriminar como ingresantes o no ingresantes a futuros postulantes a la UNALM.

VI. RECOMENDACIONES

1. Utilizar otro método de selección de individuos que beneficie mayoritariamente a las funciones discriminantes con error mínimo, de tal forma que sean las primeras al ser seleccionadas y, por consiguiente, al aplicar el cruce aritmético encontrar funciones con menor error a las iniciales.

2. Utilizar en la mutación valores aleatorios provenientes de la distribución normal, luego comparar las funciones discriminantes resultantes con las obtenidas mediante valores aleatorios de la distribución uniforme.

3. De forma computacional, realizar una prueba de desempeño a la técnica de Análisis discriminante con algoritmos genéticos, repitiéndola varias veces y verificando si se encuentran funciones discriminantes óptimas.

4. Aplicar la técnica de análisis discriminante con algoritmos genéticos no solo en el área educativa sino en otras áreas como las ciencias biológicas, ambientales, económicas, médicas, etc. como un método de optimización de la función discriminante lineal.

5. Tener en cuenta la amplia aplicación de los algoritmos genéticos en la clasificación, tomando como alternativas el análisis de regresión logística y los árboles de decisión en situaciones donde no se cumplan los supuestos básicos (Normalidad de los datos y Homogeneidad de Varianzas).

VII. REFERENCIAS BIBLIOGRÁFICAS

1. BACK, Barbro; LAITINEN, Teija; SERE, Kaise y WEZEL, Michiel. 1996. Chooosing bankruptcy predictors using discriminant analysis, logit analysis, and genetic algorithms. p. 4.

2. EFRON, B y TIBISHIRANI, R.1993. An Introduction to the bootstrap CHAPMAN. p. 239-240.

3. GESTAL, Marcos. 2010. Introducción a los Algoritmos Genéticos. Universidad de Coruña. p. 9-14.

4. GIL Natyhelem. 2006. Algoritmos Genéticos. Universidad de Colombia, Escuela de Estadística sede Medellín. p. 22.

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

6. HAIR, ANDERSON, TATHAM y BLACK. 2008. Análisis Multivariante. 5 ed. PEARSON. p. 249-279.

7. HASTIE, T; TIBSHIRANI, R y FRIEDMAN, J. 2008. The elements of statistical learning.2 ed. SPRINGER. p. 241-245.

8. HERNÁNDEZ, O. 1998. Temas de análisis estadístico multivariado. 1 ed. Universidad de Costa rica. p. 136-138

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

10. JOHNSON, Dallas. 2004. Métodos multivariados aplicados al análisis de datos. THOMSON. p. 217-274.

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

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

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

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

15. MONTANO RIVAS, Aurora; CANTÚ SIFUENTES, Mario. 2011. Algoritmos Genéticos en la discriminación. 136 p.

16. MOUJAHID, Abdelmalik; INZA, Iñaki y LARRAÑAGA, Pedro. 2008. Algoritmos Genéticos. Universidad del País Vasco. p. 1.

17. PEDRET, Ramón; SAGNIER, Laura y CAMP, Francesc. 2000. Herramientas para segmentar mercados y posicionar productos. 2 ed. DEUSTO. p. 228-234.

18. REEVES, Collin. 2010. Genetic Algorithms. School of Mathematical and Information Sciences. p. 63-64.

19. 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.

20. SHARMA, S. 1996. Applied Multivariate Techniques. WILEY. p. 263-264.

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

22. URIEL, Ezequiel y ALDÁS, Joaquín. 2005. Análisis Multivariante Aplicado. 1 ed. THOMSON. p. 278-309.

VIII. ANEXOS

ANEXO I: Funciones discriminantes de la primera iteración

Cuadro N°11. Población Inicial

n	RM	RV	MAT	FIS	QUI	BIO	(Constante)	Error	Error 1	Error 2	Error Promedio
1	2.3181	-0.2606	-0.3664	0.0508	-0.1282	0.3960	-1.3803	0.1671	0.1866	0.0576	0.1221
2	2.3061	-0.1759	-0.4153	0.1506	-0.0743	0.3974	-1.3622	0.1674	0.1882	0.0506	0.1194
3	2.4073	-0.2738	-0.3668	0.0973	-0.0768	0.3238	-1.3586	0.1750	0.1984	0.0436	0.1210
4	2.4010	-0.0854	-0.3495	0.0826	-0.0696	0.3957	-1.5131	0.1514	0.1674	0.0615	0.1145
5	2.2747	-0.1496	-0.4565	0.1616	-0.0637	0.2995	-1.2871	0.1711	0.1947	0.0384	0.1165
6	2.2668	-0.1439	-0.3546	0.1706	-0.0383	0.3043	-1.2863	0.1693	0.1925	0.0384	0.1154
7	2.3335	-0.2090	-0.4389	0.1693	-0.0533	0.4352	-1.3285	0.1761	0.1987	0.0489	0.1238
8	2.2455	-0.2251	-0.4746	0.0527	-0.0472	0.4267	-1.3007	0.1735	0.1953	0.0506	0.1229
9	2.3577	-0.1482	-0.4300	0.2139	-0.0468	0.3819	-1.3939	0.1650	0.1860	0.0471	0.1166
10	2.3250	-0.1020	-0.3916	0.2652	-0.0432	0.3587	-1.3186	0.1724	0.1956	0.0419	0.1187
11	2.3524	-0.1815	-0.3121	0.0343	-0.0111	0.4430	-1.3289	0.1700	0.1909	0.0524	0.1217
12	2.2461	-0.2674	-0.2668	0.1870	0.0240	0.4814	-1.3828	0.1648	0.1835	0.0593	0.1214
13	2.3193	-0.2232	-0.2551	0.0840	-0.1151	0.4415	-1.3684	0.1658	0.1844	0.0611	0.1228
14	2.3271	-0.2021	-0.3385	0.1899	0.0113	0.4300	-1.3743	0.1666	0.1875	0.0489	0.1182
15	2.3143	-0.1605	-0.3132	0.1885	-0.1254	0.4139	-1.3405	0.1682	0.1885	0.0541	0.1213
16	2.2587	-0.2188	-0.3877	0.2108	-0.0660	0.3403	-1.2542	0.1756	0.1999	0.0384	0.1192
17	2.3733	-0.2612	-0.3849	0.0102	-0.0162	0.4298	-1.4128	0.1645	0.1835	0.0576	0.1206
18	2.2336	-0.2134	-0.3275	0.1710	0.0329	0.3163	-1.2636	0.1740	0.1975	0.0419	0.1197
19	2.3090	-0.2257	-0.3152	0.1041	-0.1243	0.4280	-1.2924	0.1743	0.1959	0.0524	0.1241
20	2.3384	-0.2688	-0.3228	0.1928	0.0165	0.2221	-1.2968	0.1761	0.2012	0.0349	0.1180

Fuente: Elaboración Propia

Cuadro N°12. Resultados de selección

N°	ErrorPromedio	Prob Ajuste	Prob Ajuste Ac.	Selección	U_i
1	0.1221	0.0499	0.0499	Si	0.0011
2	0.1194	0.0500	0.0999	No	0.0143
3	0.1210	0.0499	0.1498	No	0.0413
4	0.1142	0.0503	0.2002	Si	0.1525
5	0.1165	0.0502	0.2503	No	0.1610
6	0.1154	0.0503	0.3006	Si	0.2584
7	0.1238	0.0498	0.3504	No	0.2862
8	0.1229	0.0498	0.4002	No	0.3178
9	0.1166	0.0502	0.4504	No	0.3361
10	0.1187	0.0501	0.5005	No	0.3506
11	0.1217	0.0499	0.5504	No	0.3540
12	0.1214	0.0499	0.6003	No	0.3757
13	0.1228	0.0498	0.6501	No	0.3992
14	0.1182	0.0501	0.7002	No	0.6030
15	0.1213	0.0499	0.7501	No	0.6883
16	0.1192	0.0500	0.8002	No	0.8078
17	0.1206	0.0500	0.8501	Si	0.8104
18	0.1197	0.0500	0.9001	No	0.8110
19	0.1241	0.0498	0.9499	No	0.8730
20	0.1180	0.0501	1.0000	Si	0.9763

Fuente: Elaboración Propia

Cuadro N°14. Funciones obtenidas mediante el Cruce Aritmético

N°	RM	RV	MAT	FIS	QUI	BIO	(Constante)	Error 1	Error 2	Error Promedio
1	2.3539	-0.1849	-0.3591	0.0645	-0.1029	0.3959	-1.4377	0.1767	0.0593	0.1180
2	2.3652	-0.1611	-0.3568	0.0689	-0.0949	0.3958	-1.4557	0.1745	0.0593	0.1169
3	2.3241	-0.2173	-0.3621	0.0721	-0.1068	0.3835	-1.3883	0.1844	0.0558	0.1201
4	2.3582	-0.1664	-0.3573	0.0720	-0.0946	0.3920	-1.4441	0.1761	0.0558	0.1160
5	2.3217	-0.1549	-0.3560	0.1130	-0.0707	0.3559	-1.3809	0.1823	0.0506	0.1164
6	2.3276	-0.1536	-0.3559	0.1083	-0.0728	0.3605	-1.3909	0.1813	0.0506	0.1160
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
9	2.3138	-0.2349	-0.3638	0.0707	-0.1117	0.3818	-1.3716	0.1866	0.0558	0.1212
10	2.3409	-0.2071	-0.3612	0.0628	-0.1091	0.3938	-1.4166	0.1798	0.0576	0.1187
11	2.3356	-0.1968	-0.3602	0.0741	-0.1008	0.3850	-1.4070	0.1801	0.0524	0.1162
12	2.2860	-0.1701	-0.3573	0.1366	-0.0622	0.3315	-1.3206	0.1888	0.0436	0.1162
13	2.3534	-0.2326	-0.3710	0.0398	-0.0743	0.4094	-1.4146	0.1823	0.0576	0.1199
14	2.3529	-0.2306	-0.3703	0.0412	-0.0766	0.4085	-1.4152	0.1823	0.0576	0.1199
15	2.3070	-0.2354	-0.3638	0.0767	-0.1088	0.3762	-1.3600	0.1875	0.0524	0.1199
16	2.2779	-0.1691	-0.3572	0.1447	-0.0577	0.3241	-1.3066	0.1906	0.0436	0.1171
17	2.3208	-0.2447	-0.3680	0.0601	-0.0962	0.3894	-1.3731	0.1882	0.0558	0.1220
18	2.3487	-0.2562	-0.3769	0.0316	-0.0582	0.4123	-1.3957	0.1866	0.0576	0.1221
19	2.3221	-0.2173	-0.3697	0.0785	-0.0477	0.3759	-1.3588	0.1882	0.0524	0.1203
20	2.3143	-0.2084	-0.3676	0.0904	-0.0488	0.3666	-1.3494	0.1888	0.0454	0.1171

Fuente: Elaboración Propia

Cuadro N°16. Funciones obtenidas mediante la Mutación Uniforme

N°	RM	RV	MAT	FIS	QUI	BIO	(Constante)	Error 1	Error 2	Error Promedio
4	2.3582	-0.1664	-0.3573	0.0720	-0.0946	0.3920	-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	2.2860	-0.1701	-0.3573	0.1366	-0.0622	0.3315	-1.4215	0.1754	0.0558	0.1156
11	2.3356	-0.1968	-0.3602	0.0741	-0.1008	0.3850	-1.4215	0.1776	0.0558	0.1167
5	2.3217	-0.1549	-0.3560	0.1130	-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
7	2.3666	-0.2053	-0.3685	0.0446	-0.0646	0.4095	-1.4215	0.1801	0.0576	0.1188
8	2.3672	-0.1996	-0.3673	0.0471	-0.0668	0.4082	-1.4215	0.1804	0.0576	0.1190
1	2.3539	-0.1849	-0.3591	0.0645	-0.1029	0.3959	-1.4215	0.1789	0.0558	0.1174

Fuente: Elaboración Propia

ANEXO II: Programas en R para la aplicación de algoritmos genéticos

2.1 Función de Cruce Aritmético

```
RCA=function(M,h,t){
g<-2*choose(h,2)
MCA<-matrix(0,g,t)
k <- 0
for(i in 1:h){
for(j in 1:h){
if(i<j){
r <- runif(1,min=0,max=1)
vy<- r*M[i,]+(1-r)*M[j,]
k <- k+1
MCA[k,]<- vy
vx<-(1-r)*M[i,]+r*M[j,]
k <- k+1
MCA[k,]<- vx
}}}
return(MCA)}
```

Argumentos:

M: Matriz de funciones

h: Número de funciones discriminantes

t: Número de variables y constante

2.2 Función de Mutación Uniforme

```
RMU=function(MUF,p1,k){
z <- sample(1:p1,1)
w <- runif(1, min(MUF[,z]), max(MUF[,z]))
MUF[,z]<-c(rep(w,k))
return(MUF)
}
```

Argumentos:

MUF: Matriz de funciones

p1: Número de funciones discriminantes

k: Número de variables y constante

2.3 Función de Evaluación

```
FEVAL=function(F,q,M,n,k){
My<-M[,1]
Mx<-M[,-1]
w<-c(rep(0,n))
y<-c(rep(1,n))
error<-c(rep(0,n))
Error_F<-c(rep(0,q))
Error1_F<-c(rep(0,q))
Error2_F<-c(rep(0,q))
ErrorPromedio<-c(rep(0,q))
Ajuste<-c(rep(0,q))
n1<-sum(My==0)
n2<-sum(My==1)
y<-c(rep(1,n))
for(i in 1:q){
for(j in 1:n){
w[j]<-sum(Mx[j,]*F[i,1:k-1])+ F[i,k]
ifelse(w[j]<=0,y[j]<-0,y[j]<-1)
ifelse(My[j]==y[j],error[j]<-1,error[j]<-0)
}
Error_F[i]<-(sum(error==0))/n
Error1_F[i]<-table(My,error)[1]/n1
Error2_F[i]<-table(My,error)[2]/n2
ErrorPromedio[i]<-(Error1_F[i]+Error2_F[i])/2
Ajuste[i]<-1-ErrorPromedio[i]
}
dataframe1<-data.frame(F, Error1_F, Error2_F, ErrorPromedio, Ajuste)
print(dataframe1)
}
```

Argumentos:

F: Matriz de funciones

q: Número de funciones discriminantes

M: Matriz de datos

n: Cantidad de datos

k: número de variables y constante

2.4 Función de Tolerancia o Filtro

```
TOL=function(F,errorf){
M<-matrix(0,dim(F)[1],dim(F)[2])
count<-0
for(i in 1:dim(F)[1]){
if(F[i,10]<=errorf)
{
count<-count+1
M<-rbind(M[1:count,],F[i,])
}
}
print(M[-1,])
}
```

Argumentos:

F: Matriz de funciones

errorf: Tolerancia

ANEXO III: Aplicación Análisis Discriminante Lineal

Se aplicará el Análisis discriminante al conjunto de datos que aparecen en el Cuadro N° 21. Este cuadro presenta datos de 20 cráneos recogidos en el Tíbet (tumbas de Sikkim y el campo de batalla de Lhasa), contiene 4 variables: la primera, Tipo de raza (1 y 2), y las tres siguientes que corresponden a medidas de Longitud, Altura de cara y Ancho de cara expresadas en mm.

Cuadro N° 21. Datos para la aplicación del Análisis Discriminante

Raza	Longitud	Altura Cara	Anchura Cara
1	190.5	73.5	136.5
1	172.5	63.0	121.0
1	167.0	69.5	119.5
1	169.5	64.5	128.0
1	175.0	77.5	135.5
1	177.5	71.5	131.0
1	179.5	70.5	134.5
1	179.5	73.5	132.5
1	173.5	70.0	133.5
1	162.5	62.0	126.0
2	195.5	78.5	144.0
2	197.0	80.5	139.0
2	182.5	68.5	136.0
2	173.5	71.5	136.5
2	188.5	79.5	136.0
2	175.0	76.5	142.0
2	196.0	76.0	134.0
2	200.0	82.5	146.0
2	185.0	81.5	137.0
2	174.5	74.0	136.5

Fuente: Técnicas de Análisis Discriminante- José R. Berrendero

Asumiendo el cumplimiento de supuestos, a continuación se presenta la función discriminante lineal de Fisher obtenida con los datos del Cuadro N° 21.

$$Z_i = -47.9529 + 0.2403 Altura\_Cara - 0.1231 Anchura\_Cara + 0.0369 Longitud \qquad (2.21)$$

La función discriminante anterior clasificó correctamente al 85% de cráneos. La clasificación correcta para los cráneos de Raza 1 fue del 80% y para los de Raza 2 del 90%.

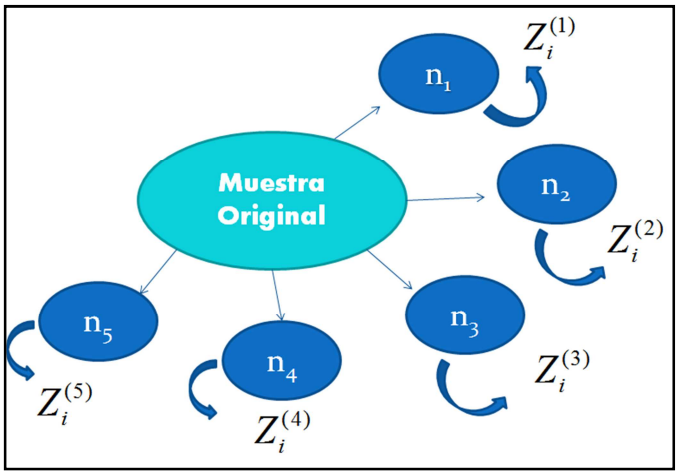
ANEXO IV: Aplicación Análisis Discriminante con Algoritmos genéticos

Se aplicará el Análisis discriminante con Algoritmos genéticos al conjunto de datos del Cuadro N° 21. Para ello, se utilizará un tamaño de población inicial igual a 5.

Paso 1: Generación de la Población Inicial

Para generar la población inicial, se aplicó un remuestreo con reemplazo a la muestra original (Cuadro N° 21). De esta manera se obtuvo 5 nuevas muestras, del mismo tamaño que la original, a las cuales se les aplicó el análisis discriminante lineal de Fisher. (Ver Gráfico N° 4).

Gráfico N° 4. Generación de la población inicial



Fuente: Elaboración propia.

Posteriormente, se toma a la muestra original y se reclasifica a los individuos con cada función discriminante generada en el paso 1. Luego se calcula el error de clasificación por grupo, tal como se muestra en el Cuadro N° 22.

Cuadro N° 22. Funciones discriminantes generadas y su error de clasificación

N°	Longitud	Alt Cara	Anch Cara	Constante	Error 1	Error 2
1	0.1739	-0.5241	0.7266	-91.0507	0.2	0
2	0.0362	0.2425	0.2388	-56.5109	0.2	0.2
3	0.0395	-0.08	0.3743	-51.5764	0.2	0
4	0.0975	-0.0336	0.3795	-66.6652	0.1	0.2
5	0.1149	-0.1451	0.4934	-76.0574	0.2	0

Fuente: Elaboración propia.

donde el Error 1 corresponde a la tasa de error de clasificación de los cráneos de la raza 1 y el Error 2 a la misma tasa pero para los cráneos de la raza 2.

Paso 2: Evaluación de la función objetivo

La función objetivo de esta aplicación, es una función discriminante lineal con menor o igual tasa de error promedio de clasificación que la obtenida por el método de Fisher (2.21), es decir con una tasa menor o igual a 0.15.

Para evaluar esta función objetivo en las demás funciones discriminantes se calcula el error promedio de clasificación de cada una de ellas y luego se ve si satisfacen o no a la función objetivo.

En el Cuadro N° 23 se puede observar que las funciones 1, 3, 4 y 5 son menores o iguales a 0.15. De las cuales sólo la 1, 3 y 5 serían soluciones factibles para el objetivo del algoritmo, pero por motivos prácticos se conservarán y se continuará con el algoritmo con la finalidad de encontrar funciones con una tasa de error promedio de clasificación aún más pequeña.

Cuadro N° 23. Funciones discriminantes y su tasa de error promedio de clasificación

N°	Longitud	Alt Cara	Anch Cara	Constante	Error 1	Error 2	Error Promedio
1	0.1739	-0.5241	0.7266	-91.0507	0.2	0	0.1
2	0.0362	0.2425	0.2388	-56.5109	0.2	0.2	0.2
3	0.0395	-0.08	0.3743	-51.5764	0.2	0	0.1
4	0.0975	-0.0336	0.3795	-66.6652	0.1	0.2	0.15
5	0.1149	-0.1451	0.4934	-76.0574	0.2	0	0.1

Fuente: Elaboración propia.

Paso 3: Cálculo de la función de aptitud

Para obtener la función de aptitud o ajuste de cada función discriminante (Cuadro N° 23) se realizará la suma de los errores promedio de clasificación, generándose el error total del problema. Por último se divide cada error promedio de clasificación entre el error total del problema, el ajuste de cada función discriminante se muestra en el Cuadro N° 24.

Cuadro N° 24. Funciones discriminantes y su ajuste

N°	Longitud	Alt Cara	Anch Cara	Constante	Error Promedio	Ajuste
1	0.1739	-0.5241	0.7266	-91.0507	0.1	0.1538
2	0.0362	0.2425	0.2388	-56.5109	0.2	0.3077
3	0.0395	-0.08	0.3743	-51.5764	0.1	0.1538
4	0.0975	-0.0336	0.3795	-66.6652	0.15	0.2308
5	0.1149	-0.1451	0.4934	-76.0574	0.1	0.1538

Fuente: Elaboración propia.

Paso 4: Selección y Operadores Genéticos

1. Selección de Individuos (Método de la ruleta)

El método de la ruleta toma los ajustes de cada función discriminante como las probabilidades de selección de cada individuo. Para ello, se generan 5 números aleatorios con distribución uniforme entre 0 y 1, uno para cada función. Luego se ordenan estos números aleatorios de menor a mayor y se comparan con la probabilidad acumulada $c_i = \sum_{j=1}^{i} P_j$. De tal forma que si un individuo i se encuentra entre $c_{i-1} < U_i(0,1) < c_i$ es seleccionado para formar la nueva población. En esta aplicación se seleccionan 3 funciones discriminantes, como se observa en el Cuadro N° 25.

Cuadro N° 25. Selección de individuos del Análisis discriminante con Algoritmos Genéticos

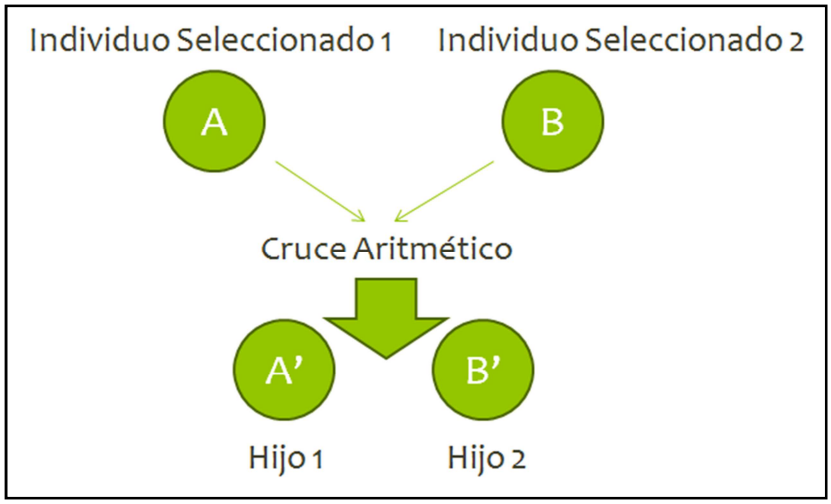
N°	Ajuste	C_{i-1}	$U_i(0,1)$	C_i	Selección
1	0.1538	0	0.3094	0.1538	No
2	0.3077	0.1538	0.428	0.4615	Si
3	0.1538	0.4615	0.5399	0.6154	Si
4	0.2308	0.6154	0.5777	0.8462	No
5	0.1538	0.8462	0.8578	1	Si

Fuente: Elaboración propia.

2. Cruce aritmético

El cruce aritmético toma dos individuos y los combina, dando como resultado de este cruce dos funciones resultantes o hijas. Una representación del cruce aritmético se observa en el Gráfico N° 5.

Gráfico N° 5. Cruce Aritmético



Fuente: Elaboración propia.

Para nuestra aplicación tomaremos las funciones discriminantes N° 2 y N° 3 del Cuadro N° 26.

Cuadro N° 26 Funciones discriminantes para la aplicación del cruce aritmético

N°	Longitud	Alt Cara	Anch Cara	Constante
2	0.0362	0.2425	0.2388	-56.5109
3	0.0395	-0.08	0.3743	-51.5764

Fuente: Elaboración propia.

Para realizar el cruce se toma en cuenta las fórmulas dadas en (2.15). Se inicia eligiendo el valor "r" de una distribución U(0,1). En este caso el valor de r fue: 0.609087. Luego se realizan los siguientes cálculos para cada columna:

Longitud:

$$S'_m = 0.609087 \times (0.0362) + (1 - 0.609087) \times (0.0395)$$
$$S'_m = 0.0375$$
$$S'_k = 0.609087 \times (0.0395) + (1 - 0.609087) \times (0.0362)$$
$$S'_k = 0.0382$$

Alt Cara:

$$S'_m = 0.609087 \times (0.02425) + (1 - 0.609087) \times (-0.08)$$
$$S'_m = 0.1164$$
$$S'_k = 0.609087 \times (-0.08) + (1 - 0.609087) \times (0.02425)$$
$$S'_k = 0.0461$$

Anch Cara:

$$S'_m = 0.609087 \times (0.2388) + (1 - 0.609087) \times (0.3743)$$
$$S'_m = 0.2918$$
$$S'_k = 0.609087 \times (0.3743) + (1 - 0.609087) \times (0.2388)$$
$$S'_k = 0.3213$$

Constante:

$$S'_m = 0.609087 \times (-56.5109) + (1 - 0.609087) \times (-51.5764)$$
$$S'_m = -54.582$$
$$S'_k = 0.609087 \times (-51.5764) + (1 - 0.609087) \times (-56.5109)$$
$$S'_k = -53.5054$$

Los dos nuevos individuos resultantes del cruce aritmético para las funciones 2 y 3 se presentan en el Cuadro N° 27:

Cuadro N° 27. Funciones discriminantes obtenidas del cruce aritmético entre las funciones N° 2 y 3

Cruce	Longitud	Alt Cara	Anch Cara	Constante
2'3	0.0375	0.1164	0.2918	-54.582
3'2	0.0382	0.0461	0.3213	-53.5054

Fuente: Elaboración propia.

Como en la aplicación se seleccionaron 3 funciones discriminantes, los pares de individuos progenitores se obtienen mediante una combinatoria de 3 en 2 (C_2^3), lo que da un resultado de 3 cruces, y como por cada cruce existen 2 funciones resultantes en total se obtendrán 6 funciones hijas ($2C_2^3$). Los resultados del cruce aritmético se muestran a continuación Cuadro N° 28.

Cuadro N° 28. Funciones discriminantes obtenidas del Cruce aritmético

N°	Cruce	Longitud	Alt Cara	Anch Cara	Constante	Error 1	Error 2	Error Promedio
1	2'3	0.0375	0.1164	0.2918	-54.582	0.2	0.1	0.15
2	3'2	0.0382	0.0461	0.3213	-53.5054	0.2	0	0.1
3	2'5	0.0436	0.2062	0.2627	-58.3451	0.2	0.2	0.2
4	5'2	0.1075	-0.1087	0.4695	-74.2233	0.2	0	0.1
5	3'5	0.0728	-0.1087	0.4268	-62.3747	0.2	0	0.1
6	5'3	0.0816	-0.1164	0.4409	-65.2591	0.2	0	0.1

Fuente: Elaboración propia.

El siguiente paso será aplicar una tolerancia a las funciones generadas en el cruce aritmético. Esta tolerancia se estableció tomando como referencia el error promedio de clasificación de la función discriminante obtenida por el método de Fisher (0.15). En consecuencia se conservarán y continuará el algoritmo con las funciones que posean un error promedio de clasificación menor o igual a 0.15, el resto se descarta.

Aplicando la tolerancia a las funciones del Cuadro N° 28, las funciones N° 1, 2, 4, 5 y 6 pasarán a la etapa de mutación del algoritmo.

3. Mutación uniforme

Para la aplicación de mutación uniforme, se elige un valor r aleatorio entre 1 y 4, valores que hacen referencia a los cuatro posibles coeficientes a mutar (Longitud, Altura de cara, Ancho de cara y constante). En este caso r es igual a 1, indicando que la columna a mutar será la correspondiente al coeficiente de Longitud de las funciones discriminantes. Posteriormente, se elige un valor aleatorio con distribución uniforme tomando como valores a y b, respectivamente, al mínimo (0.0375) y máximo (0.1075) de la columna seleccionada; resultando este valor igual a 0.0694. Luego se reemplaza todos los elementos de la columna Longitud por este valor aleatorio. Los resultados de la mutación uniforme se pueden observar en el Cuadro N° 29.

Cuadro N° 29. Funciones discriminantes obtenidas con la Mutación uniforme

N°	Longitud	Alt Cara	Anch Cara	Constante	Error 1	Error 2	Error Promedio
1	0.0694	0.1164	0.2918	-54.582	0.9	0	0.45
2	0.0694	0.0461	0.3213	-53.5054	0.9	0	0.45
4	0.0694	-0.1087	0.4695	-74.2233	0	1	0.5
5	0.0694	-0.1087	0.4268	-62.3747	0.1	0.1	0.1
6	0.0694	-0.1164	0.4409	-65.2591	0	0.6	0.3

Fuente: Elaboración propia.

Aplicando la tolerancia a las funciones del Cuadro N° 29, solo se conserva a la función número 5.

Debido a que sólo se encontró una función, el algoritmo termina y se procede a elegir la función discriminante óptima.

Paso 5: Mejores individuos y resultados

Las funciones que se conservaron en el algoritmo se presentan en el Cuadro N°30,

Cuadro N° 30. Funciones discriminantes óptimas

N°	Longitud	Alt Cara	Anch Cara	Constante	Error 1	Error 2	Error Promedio
1	0.1739	-0.5241	0.7266	-91.0507	0.2	0	0.1
2	0.0395	-0.08	0.3743	-51.5764	0.2	0	0.1
3	0.1149	-0.1451	0.4934	-76.0574	0.2	0	0.1
4	0.0694	-0.1087	0.4268	-62.3747	0.1	0.1	0.1

Fuente: Elaboración propia.

Como se puede observar en el cuadro anterior, se encontraron 4 funciones discriminantes que reportan un error promedio de clasificación de 0.1. Las tres primeras, sin error en la clasificación de cráneos de la raza tipo 2 y la última con el mismo error de clasificación tanto en las cráneos de la raza tipo 1 como 2.

Finalmente las 4 funciones discriminantes subóptimas representan soluciones factibles, ya que reportan un error promedio de clasificación menor (0.1) al obtenido aplicando sólo el método de Fisher (0.15).