RENDIMIENTO DE LOS ESTUDIANTES DEL RECINTO UNIVERSITARIO DE MAYAGÜEZ EN SU PRIMER CURSO DE MATEMÁTICAS: UNA APLICACIÓN DE ANÁLISIS MULTIVARIADO

Por:

Cidmarie E. Odiott Ruiz

Tesis sometida en cumplimiento parcial de los requisitos para el grado de

MAESTRÍA EN CIENCIAS
en
MATEMÁTICAS ESTADÍSTICAS
UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGÜEZ
Diciembre 2010

Aprobado por:

_____	_____
Edgardo Lorenzo González , Ph.D. Miembro, Comité Graduado	Fecha
_____	_____
Dámaris Santana Morant, Ph.D. Miembro, Comité Graduado	Fecha
_____	_____
Julio C. Quintana Díaz, Ph.D. Presidente, Comité Graduado	Fecha
_____	_____
Silvestre Colón Ramírez, Prof. Director Interino Departamento de Ciencias Matemáticas	Fecha
_____	_____
José R. Arroyo Caraballo, Ph.D. Representante de Estudios Graduados	Fecha

ABSTRACT

We obtained data from 6,924 freshmen admitted to the University of Puerto Rico at Mayagüez since 2005 to 2007. For each one of these students we registered information about the following variables: Type of high school (public or private), high school grade average, College Board scores examination corresponding to verbal aptitude, mathematical aptitude, mathematical knowledge, English knowledge and Spanish knowledge. College entrance index and the grade obtained in the first math course when taken by the first time. We applied two multivariate analysis: Cluster analysis to determine which variables are more related between themselves; and the discriminant analysis to obtain an inequation that would allow to identify which students would be at risk of failing their first mathematics course and which ones could be successful. We developed logistic regression model to analyze the correlation between variables, and identify those variables that are more influential on the success or failure of a student in their first mathematics course, and to obtain an equation that predicts the performance in the first year mathematics course for new admitted students to the University of Puerto Rico at Mayagüez. Comparisons were made between results of applying the discriminant analysis and logistic regression technique.

RESUMEN

Se utilizó la información de 6,924 estudiantes del Recinto Universitario de Mayagüez (RUM) que ingresaron entre los años 2005 al 2007. Para cada uno de estos estudiantes, se registraron los datos de las siguientes variables: Tipo de escuela superior (pública o privada), promedio de escuela superior, puntuaciones en las secciones del Examen de Admisión Universitaria del College Board correspondientes a la aptitud verbal, aptitud en matemáticas, aprovechamiento en matemáticas, aprovechamiento en inglés y en español. Además, su índice de ingreso a la universidad y la nota obtenida en su primer curso de matemáticas, cuando éste lo tomó por primera vez. Se aplicaron dos técnicas de análisis multivariado: un análisis por conglomerado, para determinar qué variables tienen más relación entre sí; y la técnica multivariada de análisis discriminante para obtener una desigualdad que permitiera detectar, basándose en estas variables, qué estudiantes estarían en riesgo de fracasar en su primer curso de matemáticas y cuáles podrían tener éxito en el mismo. Además, se aplicó regresión logística para analizar la correlación entre las variables, poder identificar aquellas variables que determinan el éxito o el fracaso de un estudiante en su primer curso de matemáticas, y para la obtención de un modelo que prediga el rendimiento en el primer curso de matemáticas de los estudiantes recién admitidos al RUM. Se realizaron comparaciones entre los resultados obtenidos al aplicar el análisis discriminante y la regresión logística.

DEDICATORIA

A mis padres por su apoyo y todo su amor.
A mi esposo por su cariño y comprensión.
A mis hermanas, sobrino y amigos por siempre estar presentes.

AGRADECIMIENTOS

Quiero dar gracias a Dios por nunca abandonarme en tan arduo camino y por todas las oportunidades que me ha brindado.

Quiero dar gracias al Dr. Julio C. Quintana, mi consejero, por su enseñanza de excelencia durante todos mis años de estudios, por su tiempo, paciencia, orientación y apoyo en la tesis presentada.

Quiero dar gracias a la facultad y al personal del Departamento de Ciencias Matemáticas del Recinto Universitario de Mayagüez por hacerme sentir parte de una gran familia y brindarme siempre apoyo al cursar mis estudios de Maestría.

ÍNDICE GENERAL

LISTA DE TABLAS

Página

LISTA DE FIGURAS

Página

CAPÍTULO 1
INTRODUCCIÓN

1.1 JUSTIFICACIÓN

Un por ciento considerable de los estudiantes que son admitidos al Recinto Universitario de Mayagüez (RUM) tienen dificultad en la aprobación de su primer curso de matemáticas. Las razones por las cuales estos estudiantes no logran tener éxito en su primer curso aún son inciertas, aunque puede ocurrir que factores como la escuela de procedencia, el tipo de escuela, el status socio-económico, las puntuaciones en la Prueba de Admisión Universitaria (especialmente en las áreas de matemáticas), entre otros, afecten su rendimiento. Este trabajo constituye un esfuerzo para determinar cuáles son los factores que afectan el rendimiento de los estudiantes admitidos en el RUM en su primer curso de matemáticas.

Nuestra población constó de 6,924 estudiantes que fueron admitidos en los años comprendidos entre el 2005 y el 2007 [OIIP, 2008]. Las variables a considerarse fueron el promedio de escuela superior, tipo de escuela superior (pública o privada), puntuaciones de la Prueba de Admisión Universitaria, Índice de Ingreso, Codificación de su primer curso de matemáticas, y la nota que el estudiante obtuvo en su primer curso de matemáticas.

En esta investigación se define el éxito de los estudiantes en su primer curso de matemáticas como haber obtenido una nota de A, B, C o P, y el fracaso como haber obtenido D, F, NP o no finalizar el curso (W). Nótese entonces que según la población, en el primer curso de matemáticas, un 60.75 % tuvo éxito y un 39. 25 % fracasó. La siguiente tabla muestra lo antes mencionado:

TABLA 1.1

Estadística descriptiva para el éxito y el fracaso de estudiantes

			Proporción	Por ciento
Éxito	A, B, C, P	4206	0.607	60.7
Fracaso	D, F, NP, W	2718	0.393	39.3

En esta investigación se determinará qué factores afectan significativamente el rendimiento de los estudiantes en su primer curso de matemáticas. Utilizando el análisis discriminante se establecerá un modelo matemático que se utilizará para generar una inecuación que permite predecir, con cierto margen de error, que estudiante podría éxito o fracaso en su primer curso de matemáticas. Este modelo podría utilizarse para identificar estudiantes en riesgo de fracasar y poder adoptar medidas preventivas para evitarlo. Con esta investigación se busca disminuir la proporción de estudiantes que fracasan en su primer curso de matemáticas en esta institución.

1.2 OBJETIVOS

El objetivo principal de este trabajo es determinar los factores que inciden en el éxito o en el fracaso de los estudiantes del Recinto Universitario de Mayagüez en su primer curso de matemáticas.

1.2.1 Objetivos específicos

Factores

a. Utilizando análisis discriminante y regresión logística, determinar cuáles de los datos que obtenemos al admitir a un estudiante en el RUM, tales como: el promedio de escuela superior; la escuela de procedencia; el tipo de escuela; las puntuaciones obtenidas por los estudiantes en la Prueba de Admisión Universitaria (PEAU) en las secciones de aptitud verbal, aptitud matemática, aprovechamiento en español, aprovechamiento en matemáticas y aprovechamiento en inglés; y el índice de ingreso, son factores que afectan el rendimiento de los estudiantes en su primer curso de matemáticas. Esto se logrará determinando cuáles de estas variables, según los modelos, inciden más en el éxito o fracaso de un estudiante en su primer curso de matemáticas.
b. Utilizando análisis por conglomerados, establecer las interrelaciones jerárquicas entre las variables mencionadas.

Modelo Predictivo

a. Utilizando regresión logística determinar un modelo de regresión que prediga la probabilidad de que un estudiante tenga éxito o fracaso en su primer curso de matemáticas al ser admitido al RUM.

CAPÍTULO 2
REVISIÓN DE LITERATURA

2.1 Publicaciones previas

El tema del rendimiento de los estudiantes en las matemáticas ha sido muy estudiado por años. Algunos estudios concluyen que los factores determinantes en el rendimiento de los estudiantes son la disposición del estudiante ante el aprendizaje, el tiempo que el estudiante dedica a alguna actividad de aprendizaje relacionada con la materia, la preparación académica del profesor que le dicta el curso y el estado socio-económico de la familia del estudiante. Sin embargo, no se observa consistencia en los resultados obtenidos.

En [López, Quintana, (2002)] se utilizó análisis discriminante lineal para hallar funciones utilizadas para discriminar entre dos grupos con respecto al rendimiento en los cursos de matemáticas. Los resultados de esta investigación muestran que estudiantes con buenos promedios en escuela superior y que provengan de escuelas privadas tienen mayor éxito en sus cursos de matemáticas en el Recinto Universitario de Mayagüez.

En [Castrillón, (2007)] se plantea una metodología basada en técnicas de análisis multivariado para caracterizar los estudiantes de Cálculo I (Primer semestre, 2006-07) que participaron en un estudio exploratorio. Esta caracterización se hizo con respecto a su rendimiento académico y los factores relacionados con él. Se utiliza el análisis factorial y el análisis por conglomerados en la exploración de los posibles perfiles. En esta investigación se encontró que las variables que más afectan el rendimiento del estudiante son las siguientes: 1. Nota del curso anterior. Hay una alta probabilidad de que el rendimiento futuro

sea como el rendimiento del curso previo, 2. Dificultad del examen, 3. Rapidez con la que el profesor cubrió los temas, 4. Prioridad que el estudiante le da al curso, 5. Número de veces que ha tomado el curso, 6. Número de créditos matriculados.

En [Arrieta, (1996)] se utilizó un modelo teórico jerárquico para explicar el rendimiento de estudiantes de once a doce años en sus cursos de matemáticas. Se encontró que variables como: la atención, el autoconcepto general, el cálculo y la velocidad de lectura no son significativas en el rendimiento del estudiante. En esta investigación llama la atención que las variables que tienen que ver con la familia no son relevantes en el modelo final. Las limitaciones de esta investigación es que algunos parámetros son bajos y los errores de varianza son altos. Esto es que no se mejora el porcentaje global de varianza explicada en el rendimiento en matemáticas (55%). Se obtiene un buen resultado para las variables instrumentales (del orden del 70%) y, en cambio, baja la varianza explicada en la personalidad académica, que es del orden del 33%. El modelo es mejorable en su parte derecha sobre todo en lo referente a la consideración de un mayor aporte de variables medidas en la configuración de la actitud y la personalidad académica.

En [Cervini, 2001] se utilizó el análisis estadístico por niveles múltiples para determinar cuán relacionada está la oportunidad de aprender del estudiante (OdA) y su rendimiento en el curso de matemáticas. La oportunidad de aprender está determinada por la cantidad de tiempo en clase. Se encontró que los factores que más contribuyen en el rendimiento en el curso de matemáticas son la enseñanza de los contenidos, las competencias de las pruebas y el desarrollo curricular alcanzado en general. Estos factores

son predictores relevantes en el rendimiento de un estudiante en el área de las matemáticas. Además se encontró de la OdA está fuertemente relacionada con el rendimiento de los estudiantes en sus clases de matemáticas.

En [Quintana, (2007)] se presenta un estudio descriptivo de los estudiantes de nuevo ingreso correspondientes a los años (1990-2005). En este estudio se encontró que con el pasar de los años el Índice de Ingreso (IGS) de los estudiantes aumenta significativamente a pesar que las puntuaciones en las diferentes áreas de la Prueba de Admisión Universitaria de estos estudiantes disminuyen considerablemente. Como el Índice de Ingreso es un promedio ponderado de las secciones de Aptitud Verbal y Aptitud Matemática de la PEAU y el promedio de escuela superior, entonces el incremento en el (IGS) indica que el promedio de escuela superior ha ido en aumento a pesar del descenso experimentado por las puntuaciones obtenidas por los estudiantes en los distintos componentes de la Prueba de Admisión Universitaria.

En nuestra investigación se busca determinar qué factores de los que se obtiene información en la admisión del estudiante al Recinto Universitario de Mayagüez son significativos en el éxito de su primer curso de matemáticas. Esto utilizando técnicas de análisis multivariado y modelos de regresión logística.

CAPÍTULO 3
METODOLOGÍA

En este capítulo se describe la forma en que se realizó el estudio. Se explicarán las técnicas multivariadas utilizadas en esta investigación. Se establece cuál es la población del estudio y se hace un análisis descriptivo de los mismos.

MÉTODOS DE ANÁLISIS MULTIVARIADO

3.1 Análisis por conglomerados

3.1.1 Concepto

El Análisis por Conglomerados es un conjunto de técnicas donde no se hace distinción entre variables dependientes e independientes y cuyo propósito principal es formar grupos entre las variables, estableciendo interrelaciones jerárquicas entre ellas. Tales grupos deben componerse de elementos lo más parecidos posibles entre sí y a la vez lo más diferente posible de los otros grupos [Kaufman, Rousseeuw, 1990]. Además los grupos deben ser mutuamente excluyentes, esto es, que los elementos deben pertenecer únicamente a un grupo, y ser colectivamente exhaustivos, es decir, la unión de todos los grupos debe contener a todos los elementos de la población.

3.1.2. Características

- No hay distinción entre variables dependientes e independientes.
- Grupos homogéneos entre sí y heterogéneos entre ellos.
- Permite agrupar tanto a casos o individuos pero también a variables o características.
- Es una técnica descriptiva.
- Debemos tener una muestra representativa de la población.

- No hay necesidad de los siguientes supuestos:
    - Linealidad
    - Normalidad (En nuestro caso teorema de límite central)

3.1.3. Construcción de grupos

Se requiere formar grupos de individuos lo más parecidos entre sí, por lo que utilizaremos una forma de medir el parecido de los individuos. Además debemos definir un procedimiento para definir los grupos.

Semejanza o parecido entre individuos

Para medir la semejanza o parecido entre dos individuos se utilizan medidas de similitud o distancia. Estas medidas se agrupan en tres clases:

1. Medidas de correlación

   Dos individuos son similares si tienen correlaciones altas y no serán parecidos si tienen correlaciones bajas. La correlación nos informa sobre la forma en que varían dos variables más que sobre la magnitud de las mismas.

   El coeficiente de correlación está definido por [Cuevas, Berrendero, 2003]:

$$r_{jk} = \frac{\sum \left(x_{ij} - \bar{x}_j \right)\left(x_{ik} - \bar{x}_k \right)}{\sqrt{\sum \left(x_{ij} - \bar{x}_j \right)^2 \left(x_{ik} - \bar{x}_k \right)^2}} \qquad (1)$$

   donde x_{ij} es el valor de la variable i para el caso j y $\bar{x}_j$ es la media de todos los valores para el caso j.

2. Medidas de distancia

Se consideran las magnitudes de las variables, aunque su variabilidad no tenga mucho que ver.

Algunas medidas de distancia son las siguientes:

a) Distancia Euclidiana [Cuevas, Berrendero, 2003]:

Especialmente adecuada para ejes ortogonales.

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \qquad (2)$$

donde d_{ij} es la distancia entre el caso i y j, y x_{ik} es el valor de la k^{th} variable para el i^{th} caso.

b) Distancia de Minkowski [Cuevas, Berrendero, 2003]:

$$d_{ij} = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^r \right)^{1/r} \qquad (3)$$

para r mayor o igual a 1.

c) Distancia de Mahalanobis [Cuevas, Berrendero, 2003]:

Recomendable para situaciones en las que se produce multicolinealidad.

$$d_{ij} = (X_i - X_j)' \sum^{-1} (X_i - X_j) \qquad (4)$$

Donde $\sum$ es la matriz de varianza-covarianza, y X_i y X_j son vectores de valores de las variables para los casos i y j.

3. Medidas de asociación

   Las medidas de asociación, también denominadas medidas de similitud, tienen un carácter cualitativo, se obtienen a partir de coincidencias, de acuerdos o desacuerdos.

   A mayor similitud mayor parecido entre los individuos.

Agrupación de individuos

1. Procedimientos jerárquicos

   Establecer subconjuntos disjuntos entre sí y que cada uno de ellos esté incluido en otro (jerarquía) [Dallas, 1998]. El número de grupos depende de la secuencia que consideremos. Dentro de los procedimientos jerárquicos se recurre a la representación gráfica a través de dendogramas.

   El dendograma es el tipo de gráfico más común para representar la cercanía entre los términos o variables de estudio, generando de forma automática grupos lógicos de variables, que por ejemplo, podrían formar parte de una sección o subsección [Kaufman, Rousseeuw, 1990]

   Dentro de estos procedimientos se tienen varios métodos [González, 2005]:

   a) Ascendentes

      Comienza con tantos grupos como individuos, luego se van formando grupos entre los individuos más parecidos.

      1) Vínculo único o vecino más próximo (single linkage)

      Se determina la distancia entre los objetos y luego se van agrupando objetos o individuos de acuerdo al que tenga la distancia menor.

      2) Vínculo completo (complete linkage)

      Similar al anterior solo que tomamos la distancia mayor entre ellos.

      3) Vínculo Medio (average linkage)

Se toma en cuenta la media entre los individuos. Los grupos tendrían varianzas similares y pequeñas.

4) Método del centroide

El centroide de un grupo es el punto medio en un espacio multidimensional determinado por las dimensiones o variables que se consideran en nuestro análisis. Al considerar un punto medio, los valores extraños o raros no influyen tanto en este método.

5) Método Ward

Método de la varianza porque utiliza un análisis de la varianzas para evaluar las distancias entre grupos. Se intenta minimizar la varianza intragrupo. Esto es que se forman Através de la suma total de los cuadrados de las desviaciones entre cada individuo y la media del cluster en el que se integra. Para que el proceso de clusterización resulte óptimo, en el sentido de que los grupos formados no distorsionen los datos originales, propone la siguiente estrategia:

En cada paso del análisis, considerar la posibilidad de la unión de cada par de grupos y optar por la fusión de aquellos dos grupos que menos incrementen la suma de los cuadrados de las desviaciones al unirse.

b) Descendentes

El procedimiento que se sigue es todo lo opuesto a los utilizados en los métodos ascendentes.

1) Algoritmo de Howard-Harris Utiliza el criterio de minimizar la varianza intragrupos. Es adecuado para grandes muestras.

2. Procedimientos no jerárquicos

A partir de un número de individuos, n , hay que crear k grupos siendo k un número determinado por el analista. Lo difícil de este procedimiento es establecer cuántos k grupos se deben formar. Si se crean pocos grupos se obtendrán conclusiones pobres, y

si forman demasiados grupos entonces el análisis se torna complicado [González, 2005].

Utilizando el método de análisis por conglomerados se determinará cuáles variables de las seleccionadas para la investigación se relacionan entre sí. Las variables a considerarse son: el promedio de escuela superior, las puntuaciones obtenidas por los estudiantes en las secciones de la PEAU que corresponden a aptitud verbal, aptitud matemática, aprovechamiento en español, aprovechamiento en matemáticas y aprovechamiento inglés. Además, el Índice de Ingreso y la nota del primer curso de matemáticas.

3.2. Análisis discriminante

3.2.1. Concepto

El Análisis discriminante es una técnica multivariante de dependencia, que permite encontrar funciones capaces de separar dos o más grupos de individuos tomando como base un conjunto de medidas sobre los mismos representadas por una serie de variables numéricas. Dichas funciones discriminan o identifican grupos. Es una técnica de reducción de datos [Cuevas, Berrendera, 2003]. En esta técnica multivariada se presupone la existencia de dos o más grupos bien definidos a priori.

3.2.2. Características

- La variable dependiente es categórica y las variables independientes deben ser métricas o numéricas.
- Los grupos que conforman la variable dependiente, dos o más, serán exhaustivos y mutuamente excluyentes.

- Se pretende maximizar el cociente entre la suma de los cuadrados entre grupos y la suma de los cuadrados intragrupos, con lo que se obtiene un nuevo eje que recibe el nombre de función discriminante.

3.2.3 Objetivos básicos

- Identificar las variables que mejor discriminan los grupos.
- Construir una regla de decisión (desigualdad) que asigne a uno de los grupos un individuo nuevo que no se clasificó previamente [Cuevas, Berrendero, 2003].

3.2.4. Tipos de análisis discriminante

- Análisis discriminante de dos grupos o simple: la variable dependiente tiene sólo dos categorías. Se obtiene sólo una función discriminante.
- Análisis de discriminante múltiple: la variable dependiente tiene más de dos categorías. Se puede obtener más de una función discriminante.

3.2.5. Utilidad del análisis discriminante

- Explicativos

  Con la intención de probar el poder discriminante de cada una de las variables, con la finalidad de seleccionar el subconjunto que mejor discrimina los grupos.

- Predictivos

  Se trata de encasillar a un individuo, del que no se conoce a priori a qué grupo pertenece, dentro de un grupo, a partir de los valores que toman en su caso las variables independientes.

- Reclasificadores

Ya definidos los grupos, se desea recomponer esa partición. Así muchas veces se realiza un análisis de conglomerados que posteriormente se intenta corroborar por medio de este análisis.

3.2.6. Supuestos del análisis discriminante

Para poder aplicar este análisis o para que al menos las conclusiones de este análisis sean confiables se deben cumplir los siguientes supuestos:

- La matriz de covarianzas intragrupos debe ser la misma o muy parecida en todos los grupos.
- Que la muestra siga una distribución normal multivariada.
- Ausencia de multicolinealidad entre las variables independientes.

3.2.7. Recomendaciones respecto a la muestra

- La muestra debe ser representativa de cada grupo constituido a priori.
- Las variables deben ser elegidas de manera que puedan definir y discriminar los grupos: por tanto deberían ser lo más independientes posibles unas de otras.
- Debe haber un mínimo de 20 observaciones en cada grupo.
- Se debe excluir a un individuo si no se tiene información acerca de una de las variables.

3.2.8. Funciones discriminantes

Métodos para hallar funciones discriminantes

Método directo o simultáneo: Todas las variables son incluidas simultáneamente en el análisis. Resulta adecuado cuando se desea que todas las variables intervengan en el análisis [Cuevas, Berrendero, 2003].

Método por pasos: Consiste en intentar retener el mejor conjunto de variables, entre las independientes disponibles para discriminar entre grupos. Las funciones discriminantes se pueden obtener utilizando uno de los siguientes métodos [Cuevas, Berrendero, 2003]:

- Selección hacia adelante: La variable que primero entra en el modelo es aquélla que más contribuye en la discriminación entre los grupos y de esta manera sucesivamente.
- Selección hacia atrás: En el principio se utilizan todas las variables en el modelo y luego se van descartando las variables que menos discriminen.
- Selección por pasos: Es una combinación entre los dos métodos anteriores, las variables pueden tanto entrar como salir del modelo en cualquiera de sus etapas. Las variables independientes irán entrando secuencialmente al análisis según su poder de discriminación entre los grupos.

Forma de la función discriminante

$$D = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_p x_p \qquad (5)$$

donde x_i representa los valores de cada variable independiente, a_i representa a los coeficientes estimados en base a los datos. Estos coeficientes son elegidos de tal manera que los resultados difieran lo máximo posible entre los grupos, lo cual se consigue cuando el cociente entre la suma de los cuadrados entre los grupos y la suma de los cuadrados entre grupos es máxima.

Interpretación

- Alcanzada una función discriminante con adecuada capacidad de clasificación, cabe la posibilidad de identificar perfiles de los individuos que forman los diferentes grupos.
- Para la interpretación de la función hay tres indicadores de valiosa ayuda:
    - Coeficientes estandarizados de la función discriminante

      La interpretación de los coeficientes de la función discriminante es similar a la que se hace en el análisis de regresión múltiple. Los signos de los coeficientes son arbitrarios, pero van a incidir en un mayor o menor valor de la función y se podrán asociar a determinados grupos.
    - Estructura de correlaciones

      Su interpretación es equivalente a la que se haría en un análisis de varianzas en el que los resultados discriminantes fuesen la variable dependiente. Cuanto mayor sea la correlación, más importante sería el papel de la variable en la función discriminante.
    - Valores parciales de la F para la variable

      Un valor elevado de F está asociado con un alto poder discriminante.

3.2.9 Tasa de mala clasificación

La tasa de mala clasificación es la probabilidad de que el clasificador clasifique mal una observación de la población a la cual pertenece la muestra usada para construir el clasificador.

La técnica de análisis multivariado será utilizada para la discriminación entre los grupos de éxito o fracaso de un estudiante en su primer curso de matemáticas. Se discriminará utilizando las variables que se mencionaron previamente en el análisis por conglomerados, con la excepción del tipo de escuela, que es una variable categórica.

3.3. Regresión Logística

3.3.1 Concepto

La regresión logística [Hosmer, Lemeshow, 2000] es una técnica de análisis multivariado en el cual se intenta predecir si un determinado suceso ocurrirá o no en función de un conjunto de variables explicativas. Se trata de construir un modelo que describa la relación entre una serie de características que forman un conjunto de variables independientes de tipo categórico o continuo y una variable dependiente que puede ser binaria o dicotómica que sólo puede tomar dos valores que definen opciones o características opuestas o mutuamente excluyentes.

El análisis discriminante servirá para abordar situaciones como las descritas; sin embargo, la posibilidad de que coexistan variables independientes de naturaleza cuantitativa y categórica viola el supuesto de normalidad multivariante. El modelo de regresión logística es un procedimiento por medio del cual se intenta analizar las relaciones de asociación entre una variable dependiente Y y una o más variables independientes (regresores o predictores) X_n cuantitativas o categóricas, todo esto con el fin de lograr los siguientes objetivos:

- Determinar la existencia o ausencia de la relación entre una o más variables independientes y la variable dependiente.
- Medir la magnitud de dicha relación y estimar o predecir la probabilidad de que se produzca o no el suceso definido por la variable dependiente en función de los valores que adopten las variables independientes.

3.3.2 Modelo de Regresión Logística

Una excelente alternativa para garantizar que la respuesta prevista esté entre 0 y 1 es utilizar una función de enlace no lineal que sea monótona, creciente y acotada entre dichos valores (0 y 1). En estas circunstancias cabrá utilizar cualquier función de distribución de variables aleatorias, de tal modo que el modelo quedaría: $p_i = F(\beta_0 + \beta_1 x_i)$ es decir, que la probabilidad de que ocurra el evento descrito por la variable dependiente es $P(Y = 1/i) = p_i$ viene expresada por una función de distribución (no lineal) de sus variables independientes (x_i). El modelo de regresión logística simple surge al utilizarse la función de distribución logística para modelar la relación entre la probabilidad de $Y = 1$, condicionada a un determinado valor de las variables independientes, x_i

$$p_i = \frac{e^{\beta_1 + \beta_2 x_i}}{1 + e^{\beta_1 + \beta_2 x_i}} \tag{6}$$

$$(1 - p_i) = \frac{1}{1 + e^{\beta_1 + \beta_2 x_i}} \tag{7}$$

Para el ajuste de este modelo y la estimación de los parámetros $\hat{\beta}_1$ y $\hat{\beta}_2$ no puede seguirse el método de mínimos cuadrados puesto que, como se ha comentado, cuando se aplica al caso

de variables dependientes dicotómicas, el modelo resultante presenta heteroscedasticidad. Una alternativa de uso general para la estimación de los parámetros consiste en utilizar el procedimiento de estimación por máxima verosimilitud (EMV) [Draper, Smith, 1998]. Este método proporciona unos valores ($\hat{\beta}_1$ y $\hat{\beta}_2$) para los parámetros desconocidos (β_1 y β_2) que minimizan la probabilidad de que con ellos se obtengan los valores observados. Así pues una vez obtenidos los estimadores de máxima verosimilitud $\hat{\beta}_1$ y $\hat{\beta}_2$, la estimación de la probabilidad $\hat{p}_2$ es inmediata:

$$\hat{p}_2 = \frac{e^{\hat{\beta}_1+\hat{\beta}_2 x_2}}{1+e^{\hat{\beta}_1+\hat{\beta}_2 x_2}} \tag{8}$$

$$\left\langle - \hat{p}_2 \right\rangle = \frac{1}{1+e^{\hat{\beta}_1+\hat{\beta}_2 x_2}} \tag{9}$$

En el caso donde se utilicen k variables independientes, la ecuación de regresión logística que se utilizaría seria:

$$\hat{p}_k = \frac{e^{\hat{\beta}_1+\hat{\beta}_2 x_2+\ldots+\hat{\beta}_k x_k}}{1+e^{\hat{\beta}_1+\hat{\beta}_2 x_2+\ldots+\hat{\beta}_k x_k}} \tag{10}$$

$$\left\langle - \hat{p}_k \right\rangle = \frac{1}{1+e^{\hat{\beta}_1+\hat{\beta}_2 x_2+\ldots+\hat{\beta}_k x_k}} \tag{11}$$

POBLACIÓN DEL ESTUDIO

3.4 Recolección de datos

Los datos consisten de 6,924 estudiantes de primer ingreso al Recinto Universitario de Mayagüez, entre los años 2005 al 2007, con las siguientes variables: Tipo de escuela

superior (pública o privada), promedio de escuela superior, puntuaciones en las secciones de aptitud verbal, aptitud en matemáticas, aprovechamiento en matemáticas, aprovechamiento en inglés y en español, su índice de ingreso a la universidad y la nota obtenida en su primer curso de matemáticas, cuando éste lo tomó por primera vez. Estos datos fueron provistos por la Oficina de Investigación Institucional y Planificación (OIIP) del RUM.

3.5 Perfil de la población del estudio

La población consiste de 6,924 estudiantes matriculados en su primer curso de matemáticas. Los cursos que estos estudiantes tomaron fueron: MATE 0066, MATE 3005, MATE 3018, MATE 3021, MATE 3023, MATE 3031, MATE 3049, MATE 3086, MATE 3143, MATE 3151, MATE 3171 y MATE 3172.

La Tabla 3.1 nos muestra los cursos de matemáticas con su codificación, su respectivo título y número de horas contacto a la semana o créditos del mismo.

TABLA 3.1
Codificación, título y número de créditos de los cursos de matemáticas

Codificación	Curso	Número de horas/crédito
MATE 0066	Matemática Prebásica	Tres horas sin crédito
MATE 3005	Precálculo	Cinco
MATE 3018	Precálculo y Geometría Analítica	Cuatro
MATE 3021	Cálculo para las Ciencias Biológicas I	Tres
MATE 3023	Precálculo I	Dos
MATE 3031	Cálculo I	Cuatro
MATE 3049	Análisis Matemático para las Ciencias Gerenciales	Tres
MATE 3086	Razonamiento Matemático	Tres

MATE 3143	Cálculo con precálculo I	Cinco
MATE 3151	Cálculo I	Cuatro
MATE 3171	Precálculo I	Tres
MATE 3172	Precálculo II	Tres

De los 6, 924 estudiantes, 4,206 (60.7 %) lograron tener éxito en su primer curso de matemáticas y 2,718 (39.3 %) fracasaron en el mismo. De éstos el 3,903 (56.4 %) eran provenientes de escuelas públicas y 3021 (43.6 %) provenían de escuelas privadas. Del 43.6 % de los estudiantes que provinieron de las escuelas privadas el 65.1% tuvo éxito en su primer curso y que del 56.4 % de los estudiantes que provinieron de las escuelas públicas el 57.4 % de los estudiantes lograron tener éxito.

La siguiente tabla muestra las medias de las variables numéricas para el grupo completo.

Tabla 3.2
Media de las variables numéricas para el grupo completo

Variable	Media
Promedio de escuela superior	3.65
Aptitud verbal	585
Aptitud matemática	629
Aprovechamiento en inglés	579
Aprovechamiento matemático	625
Aprovechamiento en español	543
IGS	317

La siguiente tabla muestra las medias de las variables numéricas por tipo de escuela (Pública o privada).

Tabla 3.3
Medias para las variables por tipo de escuela superior (pública o privada)

Variable	Tipo de Escuela	Media
Promedio de escuela superior	Privada	3.58
	Pública	3.70
Aptitud verbal	Privada	600.84
	Pública	572.06
Aptitud matemática	Privada	649.72
	Pública	612.69
Aprovechamiento en inglés	Privada	625.78
	Pública	542.34
Aprovechamiento matemático	Privada	647.61
	Pública	606.76
Aprovechamiento en español	Privada	555.90
	Pública	532.98
IGS	Privada	319.84
	Pública	314.66

En la tabla 3.3 podemos observar cómo afecta el tipo de escuela a las diferentes variables del estudio. Notamos que los estudiantes procedentes de las escuelas privadas obtienen los índices de ingreso universitario y puntuaciones en la Prueba de Admisión Universitaria del College Board más altos que los de las escuelas públicas, sin embargo obtienen promedios de escuela superior más bajos que los estudiantes procedentes de las escuelas públicas. Esto significa que los estudiantes de escuelas públicas ingresan con promedios de escuela más

altos, mientras que sus puntuaciones en la Prueba de Admisión Universitaria del College Board son más bajas.

CAPÍTULO 4
RESULTADOS

En este capítulo se explicará el análisis de la información que maneja la Oficina de Investigación Institucional y Planificación para calcular la similitud entre las variables. Además se calculará el modelo matemático para discriminar entre el éxito y el fracaso de los estudiantes en el primer curso de matemáticas.

4.1 Análisis de conglomerados

En el análisis por conglomerados se consideraron las siguientes variables:

PROGRAM_ESTUDIO_ADMITIDO: Programa de estudio al que el estudiante fue admitido

PROMEDIO_ESC_SUP: Promedio de escuela superior

VERBAL_APT: Puntuación de Aptitud Verbal en la Prueba de Admisión Universitaria

MATE_APT: Puntuación de Aptitud Matemática en la Prueba de Admisión Universitaria

MATE_APR: Puntuación de Aprovechamiento en Matemáticas en la Prueba de Admisión Universitaria

INDICE_DE_INGRESO: Índice de ingreso a la universidad

ESPA_APR: Puntuación de Aprovechamiento en Español en la Prueba de Admisión Universitaria

INGL_APR: Puntuación de Aprovechamiento en Inglés en la Prueba de Admisión Universitaria

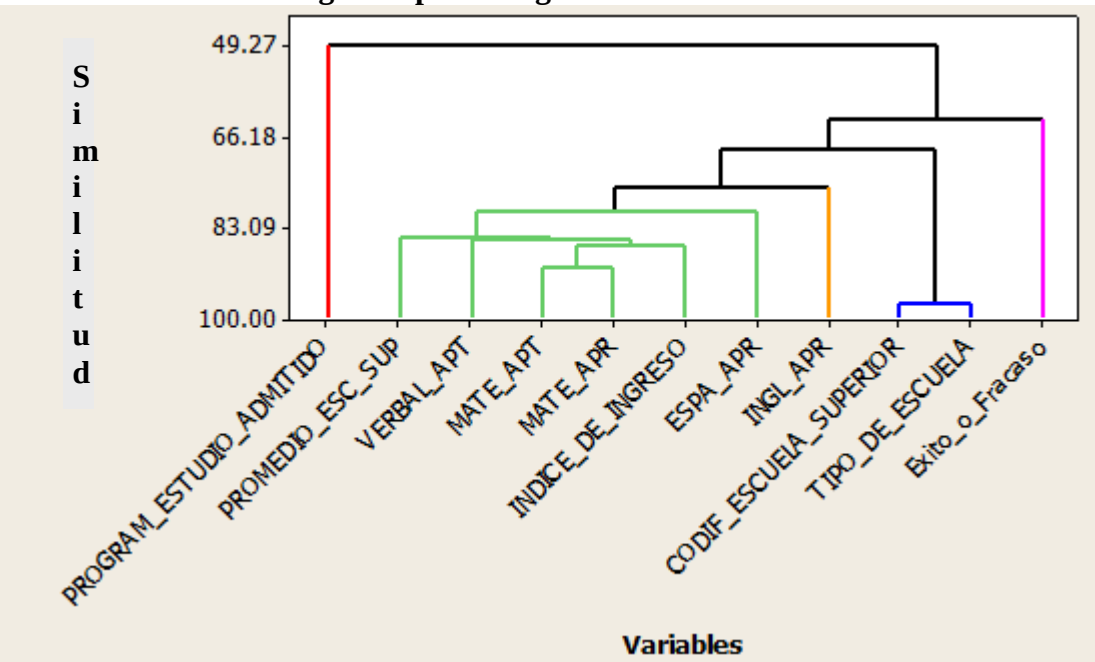
CODIF_ESCUELA_SUPERIOR: Código de la escuela superior de procedencia

TIPO_DE_ESCUELA: Tipo de Escuela de procedencia: pública o privada

ÉXITO O FRACASO: Fracaso = 0, Éxito = 1

La Figura 4.1 nos muestra la similitud entre las siguientes variables: programa de estudio al que fue admitido el estudiante, promedio de escuela superior, puntuaciones en la Prueba de Admisión Universitaria del College Board, índice de admisión universitaria (IGS), escuela de procedencia, tipo de escuela de procedencia y el éxito o el fracaso de los estudiantes en su primer curso de matemáticas.

FIGURA 4.1
Dendograma para conglomerados de variables

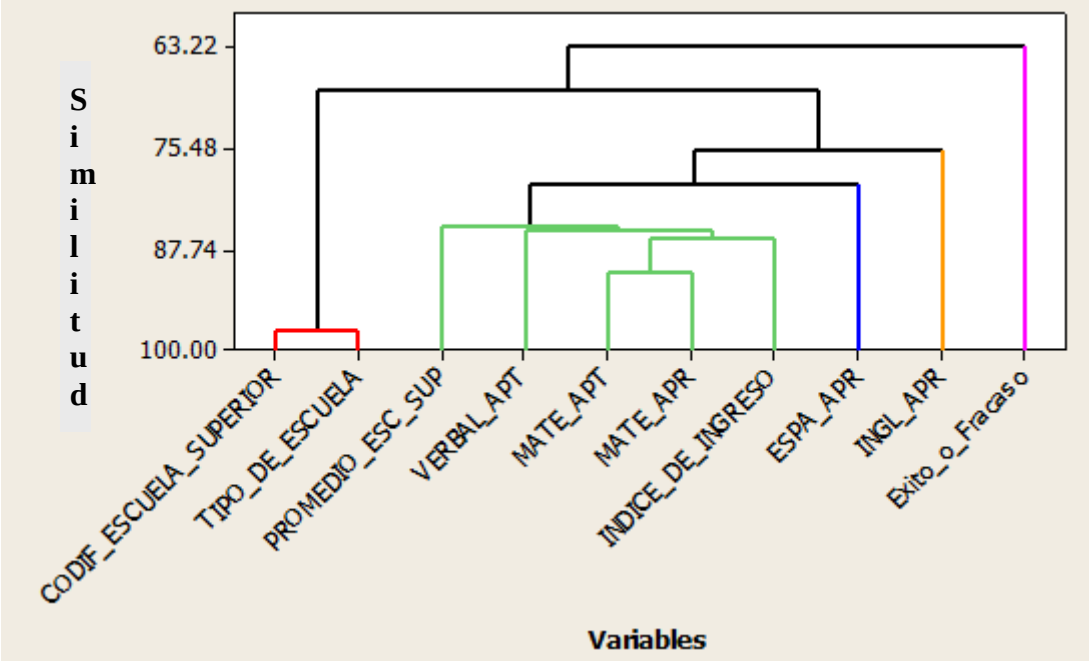


De la Figura 4.1 se puede afirmar que:

- El código de la escuela y el tipo de escuela de procedencia son 97.56 % similares.
- La puntuación en la prueba de aprovechamiento matemático y la puntuación en aptitud matemática son 90.55 % similares.
- Las puntuaciones en las áreas de matemáticas en la Prueba de Admisión Universitaria y el Índice de Ingreso son 86. 41 % similares.
- El éxito o fracaso se relaciona con todas las variables excepto con el programa de estudio al que el estudiante fue admitido.
- Podemos entonces identificar a los estudiantes en riesgo de fracasar sin tener en cuenta el programa de estudio al que el estudiante fue admitido, esto posiblemente nos puede encaminar a poder decir, con más estudios, que la carga académica y los otros cursos que los estudiantes toman a la vez que su primer curso de matemática no afecta el rendimiento de estos en los mismos.

La siguiente figura nos muestra la similitud entre las siguientes variables: promedio de escuela superior, puntuaciones en la Prueba de Admisión Universitaria del College Board, índice de admisión universitaria (IGS), Escuela de procedencia, tipo de escuela de procedencia y el éxito o el fracaso de los estudiantes en su primer curso de matemáticas. La diferencia con el anterior es que el siguiente dendograma excluye el programa de estudio al que el estudiante fue admitido.

FIGURA 4.2
Dendograma para conglomerados de variables



De la Figura 4.2 podemos afirmar:

- La similitud y relación entre el éxito o el fracaso de un estudiante en su primer curso de matemáticas y las variables antes mencionadas es de un 63.22 %.
- En el segundo nivel de similitud o de relación con las variables está el tipo y el código de la escuela, con las variables restantes, con un 68.65 % de similitud.
- En el tercer nivel de similitud están el Aprovechamiento en Inglés y el Aprovechamiento en Matemáticas, con un 75.84% de similitud y las variables restantes.
- En el cuarto nivel de similitud se encuentra la puntuación de aprovechamiento en español con un 80.05 % de similitud con las variables restantes.

- En el quinto nivel de similitud se encuentra el resto de las variables con un 85.44 % de similitud.

4.2 Análisis discriminante

Utilizando el análisis discriminante se obtiene un modelo descrito de la siguiente manera:

Éxito_Fracaso = f(Promedio de Escuela, Verbal Apt, Mate Apt, Inglés Apr, Mate Apr, Espa Apr)

Descripción de cada Variable:

Éxito o fracaso: Fracaso = 0, Éxito = 1

Promedio de Escuela: Promedio de escuela superior

Verbal Apt: Puntuación de Aptitud verbal en la Prueba de Admisión Universitaria

Mate Apt: Puntuación de Aptitud matemática en la Prueba de Admisión Universitaria

Inglés Apr: Puntuación de Aprovechamiento académico en ingles en la Prueba de Admisión Universitaria

Mate Apr: Puntuación de Aprovechamiento académico en matemáticas en la Prueba de Admisión Universitaria

Espa Apr: Puntuación de Aprovechamiento académico en español en la Prueba de Admisión Universitaria

En esta sección se utilizará la técnica estadística multivariante de Análisis Discriminante para el modelo final del cálculo del éxito o fracaso de un estudiante en el primer curso de matemáticas en el Recinto Universitario de Mayagüez. La variable respuesta es categórica

(éxito o fracaso) y toma el valor de cero cuando se trata del fracaso del estudiante y toma el valor uno cuando el estudiante tiene éxito en su primer curso.

TABLA 4.1
Estadísticos descriptivos para éxito o fracaso de los estudiantes en su primer curso de matemáticas

Variable	Rendimiento	Media	Diferencias
Promedio de escuela superior	Éxito	3.72	0.18
	Fracaso	3.54	
Aptitud verbal	Éxito	590.79	15.73
	Fracaso	575.06	
Aptitud matemática	Éxito	642.25	**33.75**
	Fracaso	608.50	
Aprovechamiento en inglés	Éxito	584.88	15.63
	Fracaso	569.25	
Aprovechamiento matemático	Éxito	638.50	**35.46**
	Fracaso	603.04	
Aprovechamiento en español	Éxito	550.63	19.49
	Fracaso	531.14	
IGS	Éxito	323.53	16.83
	Fracaso	306.70	

De la Tabla 4.1 se puede observar que todos los promedios de las variables para el Grupo que tuvo éxito en su primer curso de matemáticas, son más altos que los del Grupo que fracasó en su primer curso de matemáticas. Notamos además que las diferencias más significativas en lo que concierne a las Pruebas de Admisión y Aprovechamiento del College Boad están entre las siguientes variables: puntuaciones en Aptitud matemática y aprovechamiento matemático.

TABLA 4.2
Matriz de Correlación de Variables (Grupo Completo)

	Promedio de Escuela	Verbal Apt	Mate Apt	Inglés Apr	Mate Apr	Espa Apr	IGS
Promedio de Escuela	1.000						
Verbal Apt	.170	1.000					
Mate Apt	.108	.467	1.000				
Ingles Apr	.026	.514	.448	1.000			
Mate Apr	.171	.489	**.804**	.498	1.000		
Espa Apr	.235	.596	.372	.434	.422	1.000	
IGS	**.680**	**.710**	**.716**	.428	**.668**	.542	1.000

De la tabla 4.2 podemos afirmar que las variables más correlacionadas son las siguientes: IGS se relaciona con las siguientes variables: promedio de escuela superior, puntuación de aptitud matemática, puntuación de aptitud matemática y con el aprovechamiento matemático. Esto tiene mucho sentido pues para calcular el IGS se toma en cuenta el promedio de escuela superior, la puntuación de aptitud verbal y la puntuación de aptitud matemática. También tenemos que la aptitud matemática tiene una alta correlación con el aprovechamiento matemático de los estudiantes.

Además debemos resaltar que el promedio de escuela superior tiene correlaciones muy bajas con todas las variables, excepto con el IGS. Esto podría sugerir que las altas correlaciones entre el IGS y las variables antes mencionadas se deben más al componente de la PEAU que tiene el IGS que al promedio de escuela superior.

La tabla 4.3 lleva a establecer la relación entre las variables explicativas y la función discriminante, con el fin de determinar cuáles son las variables fundamentales en el modelo

discriminante. Esto se analiza a través de los coeficientes estandarizados, ignorando su signo. El valor absoluto de cada coeficiente asociado a cada variable determinará el grado de contribución de esta variable a la discriminación. A mayor absoluto de un coeficiente, mayor efecto de la variable en la discriminación. Los que más contribuyen y no se pueden extraer del modelo son las siguientes: el promedio de escuela superior y la puntuación en aptitud matemática.

TABLA 4.3
Coeficientes estandarizados para la función discriminante

Promedio de Escuela	**.741**
Verbal Apt	-.085
Mate Apt	**.435**
Ingles Apr	-.081
Mate Apr	.198
Espa Apr	.090

La tabla 4.4 representa la correlación lineal entre cada variable y la función discriminante, aquellas variables que tienen coeficientes de correlación altos son las variables que más aportan a la función discriminante.

Tabla 4.4
Matriz de estructura en análisis discriminante

Promedio de Escuela	**.826**
Mate Apt	**.631**
Mate Apr	**.630**
Espa Apr	.423
Verbal Apt	.352
Inglés Apr	.226

De las tablas 4.3 y 4.4 se puede concluir que el promedio de escuela superior, la puntuación de aprovechamiento matemático y la puntuación de aptitud matemática son las variables que más contribuyen a discriminar entre los grupos.

A continuación centraremos el análisis en verificar si discrimina más a los grupos éxito o fracaso en el primer curso de matemática; la puntuación en aptitud verbal, la puntuación en aptitud matemática, la puntuación en aprovechamiento en ingles, la puntuación en aprovechamiento matemático, la puntuación en aprovechamiento en español o el promedio de escuela superior.

La prueba Lambda de Wilks es usada para determinar la significancia de las variables que se utilizan para discriminar. [Dallas, 1998] El estadístico Lambda de Wilks se obtiene al calcular la razón entre el determinante de la matriz de variancias y covariancias dentro de grupos y el determinante de la matriz de variancias y covariancias total; y puede tener una aproximación asintótica con la distribución F.

Hipótesis:

$H_0 =$ la variable no es significativa para discriminar entre los grupos
$H_1 =$ la variable es significativa para discriminar entre los grupos

TABLA 4.5
Prueba Lambda de Wilks de significatividad discriminante

	Wilks' Lambda	F	df1	df2	Significatividad
Verbal Apt	.989	79.640	1	6922	**.000**
Mate Apt	.964	256.038	1	6922	**.000**
Inglés Apr	.995	32.925	1	6922	**.000**
Mate Apr	.964	255.029	1	6922	**.000**
Espa Apr	.984	114.893	1	6922	**.000**
Promedio de Escuela	.940	438.399	1	6922	**.000**

Tomando un alfa = 0.05 los datos de la Tabla 4.5 ponen de relieve que todas las variables antes mencionadas discriminan significativamente y deben ser utilizadas en el modelo discriminante para el éxito o el fracaso de los estudiantes en su primer curso de matemáticas.

La Tabla 4.6 muestra la prueba Lambda de Wilks para la función discriminante. Esta prueba se utiliza para la validación del modelo.

Hipótesis:

$H_0 =$ El modelo no es válido

$H_1 =$ El modelo es válido

TABLA 4.6
Validación del Modelo
Wilks' Lambda

Prueba para la función discriminante	Wilks' Lambda	Chi-square	Grados de libertad	p
Éxito o fracaso	.915	614.751	6	.000

Utilizando la prueba Lambda de Wilks y un alfa = 0.05 obtenemos la validación del modelo discriminante pues el valor p del modelo es 0.000 < 0.05., lo que nos indica que se rechaza la hipótesis nula H_0.Podemos concluir que el modelo es válido.

La Tabla 4.7 nos muestra los coeficientes no estandarizados para la función o modelo discriminante.

TABLA 4.7
Coeficientes de la función discriminante

x_1	Promedio de Escuela	2.174
x_2	Verbal Apt	-.001
x_3	Mate Apt	.005
x_4	Ingles Apr	-.001
x_5	Mate Apr	.002
x_6	Espa Apr	.001
	Constante	-12.027

Entonces tenemos que la función discriminante es:

$$Y = -12.027 + 2.174x_1 - 0.001x_2 + 0.005x_3 - 0.001x_4 + 0.002x_5 + 0.001x_6 \qquad (12)$$

TABLA 4.8
Clasificación de resultados por error aparente

	Asignados por el modelo			
Datos Reales		0 Fracaso	1 Éxito	Totales
	0 Fracaso	1568	1150	2718 (39.3%)
	1 Éxito	1347	2859	4206 (60.7 %)
	Totales	2915 (42.1 %)	4009 (57.9 %)	6924 (100 %)

La Tabla 4.8 muestra las clasificaciones originales y las clasificaciones hechas por la función discriminante. Nótese que el 63.9 % de los estudiantes de la población fueron correctamente clasificados. Además se puede observar que en el grupo 0, grupo que representa el fracaso, 1,150 estudiantes fracasaron y sin embargo de acuerdo al modelo debieran ser clasificados como éxitos, es decir, un error de 16.6% por ese tipo de mala clasificación. En el grupo 1,

grupo que representa el éxito, 1,347 estudiantes que tuvieron éxito serían de acuerdo al modelo clasificados erróneamente como fracasos, lo que representa un error de 19.5% por este otro tipo de mala clasificación. Por lo tanto, la tasa total de mala clasificación fue de 0.361, es decir que el $2497/6924 = 36.1\%$ de los datos fueron clasificados incorrectamente por el modelo.

Un falso positivo ocurre cuando se predice éxito del estudiante y en realidad el estudiante fracasa. Nuestra regla de decisión predijo que 4,009 de los estudiantes tendría éxito. Esta predicción fue errónea para 1,150 de los estudiantes, por lo que la tasa del falso positivo es $1150/4009 = 28.7\%$.

Un falso negativo ocurre cuando se predice el fracaso del estudiante y en realidad el estudiante obtiene éxito. Nuestra regla de decisión predijo que 2,915 de los estudiantes fracasarían en su primer curso de matemáticas. Esta predicción fue errónea para 1,347 de los estudiantes, por lo que la tasa del falso negativo es $1347/2915 = 46.2\%$.

4.3 Regresión logística para variables numéricas

En esta sección se utilizará la técnica multivariada de regresión logística para la obtención un modelo final para el cálculo del éxito o el fracaso de un estudiante en su primer curso de matemáticas. En este modelo la variable de respuesta es la variable dicotómica que toma valor 0 si el estudiante fracasa, es decir, obtiene una nota de D, F, NP o W en su primer curso de matemáticas o un valor de 1 si el estudiante tiene éxito, es decir, obtiene una nota de A, B, C o P, en su primer curso de matemáticas. Para poder realizar comparaciones con la

técnica del Análisis Discriminante, no se consideró en el modelo de regresión logística la variable Tipo de Escuela, la cual se eliminó al aplicar dicha técnica, por ser categórica.

El modelo matemático es el siguiente:

Éxito_Fracaso = f(Promedio de Escuela,Verbal Apt, Mate Apt, Inglés Apr, Mate Apr, Espa Apr)

Descripción de cada Variable:

Éxito o fracaso: Fracaso = 0, Éxito = 1

Promedio de Escuela: Promedio de escuela superior

Verbal Apt: Puntuación de Aptitud verbal en la Prueba de Admisión Universitaria

Mate Apt: Puntuación de Aptitud matemática en la Prueba de Admisión Universitaria

Inglés Apr: Puntuación de Aprovechamiento académico en ingles en la Prueba de Admisión Universitaria

Mate Apr: Puntuación de Aprovechamiento académico en matemáticas en la Prueba de Admisión Universitaria

Espa Apr: Puntuación de Aprovechamiento académico en español en la Prueba de Admisión Universitaria

La Tabla 4.9 muestra los coeficientes de la función de regresión logística para cada una de las variables independientes en el modelo.

TABLA 4.9
Coeficientes para la función de regresión logística

x_1	PROMEDIO_ESC_SUP	1.291
x_2	VERBAL_APT	-.001
x_3	MATE_APT	.003
x_4	INGL_APR	.000
x_5	MATE_APR	.001
x_6	ESPA_APR	.001
	Constante	-6.741

Por lo tanto, la función estimada de regresión logística para el éxito o fracaso de los estudiantes es la siguiente:

$$p_i = \frac{e^{-6.741+1.291x_1-0.001x_2+0.003x_3+0.001x_5+0.001x_6}}{1+e^{-6.741+1.291x_1-0.001x_2+0.003x_3+0.001x_5+0.001x_6}} \qquad (13)$$

donde p_i es la probabilidad de que un estudiante obtenga un valor 0 ó un valor 1 al sustituir sus datos correspondientes a las variables independientes en la ecuación anterior.

TABLA 4.10
Correlación entre variables del modelo de regresión (Grupo de fracasos)

	PROMEDIO_ ESC_SUP	VERBAL _APT	MATE_ APT	INGL_ APR	MATE_ APR	ESPA_ APR
PROMEDIO_ESC_SUP	1.000					
VERBAL_APT	**-.062**	1.000				
MATE_APT	**.067**	-.116	1.000			
INGL_APR	**.133**	-.258	-.040	1.000		
MATE_APR	**-.114**	-.067	**-.722**	-.188	1.000	
ESPA_APR	**-.167**	-.420	.007	-.154	-.079	1.000

La tabla 4.10 muestra la correlación que existe entre las variables del modelo de regresión logística para el grupo de estudiantes que fracasó en su primer curso de matemáticas. Se puede observar que las puntaciones en Aptitud Matemática de la PEAU y las puntuaciones de Aprovechamiento Matemático están altamente correlacionadas negativamente. Lo más relevante que podemos observar es que no hay correlación entre la significativa entre los promedios de escuela superior y el conocimiento medido por la Prueba de Admisión Universitaria. Notamos además correlaciones negativas entre el promedio de escuela superior y las puntuaciones de aptitud verbal, aprovechamiento matemático y aprovechamiento en español, esto indica que mientras alguna de estas variables aumenta la otra disminuye, esto es que si el promedio de estos estudiantes aumenta su rendimiento en las áreas antes mencionadas disminuye.

TABLA 4.11
Correlación entre variables del modelo de regresión (Grupo de éxitos)

	PROMEDIO_ ESC_SUP	VERBAL _APT	MATE_ APT	INGL_ APR	MATE_A PR	ESPA_ APR
PROMEDIO_ESC_SUP	1.000					
VERBAL_APT	**.213**	1.000				
MATE_APT	**.134**	.490	1.000			
INGL_APR	**.052**	.519	.443	1.000		
MATE_APR	**.199**	.501	**.788**	.494	1.000	
ESPA_APR	**.257**	.593	.379	.438	.414	1.000

La tabla 4.11 muestra la correlación que existe entre las variables del modelo de regresión logística para el grupo de estudiantes que obtuvo éxito en su primer curso de matemáticas. Se puede observar las bajas correlaciones entre las puntuaciones en las diferentes partes de la

Prueba de Admisión Universitaria, esto es que no hay correlación entre la significativa entre los promedios de escuela superior y el conocimiento medido por la Prueba de Admisión Universitaria.

Al comparar estas dos matrices de correlación tenemos que en el grupo de fracasos se observa la alta correlación negativa entre las puntuaciones en las secciones de aptitud matemática y aprovechamiento matemático. Sin embargo se observa lo opuesto para el grupo de éxitos una alta correlación positiva entre las secciones de aptitud matemática y aprovechamiento matemático.

Reglas de decisión para clasificar a los estudiantes de nuevo ingreso en cuanto a su posibilidad de tener éxito o fracaso en su primer curso de matemáticas en el RUM:

Si $\frac{e^{-6.741+1.291x_1-0.001x_2+0.003x_3+0.001x_5+0.001x_6}}{1+e^{-6.741+1.291x_1-0.001x_2+0.003x_3+0.001x_5+0.001x_6}} \geq 0.62$, entonces se considera que el estudiante tendrá éxito; y si $\frac{e^{-6.741+1.291x_1-0.001x_2+0.003x_3+0.001x_5+0.001x_6}}{1+e^{-6.741+1.291x_1-0.001x_2+0.003x_3+0.001x_5+0.001x_6}} < 0.62$, entonces se considera que el estudiante fracasará.

Para clasificar cada estudiante, cada dato, en el éxito o fracaso en su primer curso de matemáticas utilizamos las desigualdades antes mencionadas. Si al evaluar cada variable de un caso se obtiene un número menor que 0.62 se espera que el estudiante no tenga éxito. Por el contrario, si evaluamos y obtenemos un número mayor o igual que 0.62, entonces este estudiante se espera que culmine con éxito el primer curso de matemáticas.

TABLA 4.12
Clasificación de resultados por error aparente

	Asignados por el modelo			
Datos Reales		0 Fracaso	1 Éxito	Totales
	0 Fracaso	1674	1044	2718 (39.3 %)
	1 Éxito	1500	2706	4206 (60.7 %)
	Totales	3174 (45.8 %)	3750 (54.2 %)	6924 (100 %)

La tabla 4.12 muestra el número de datos en cada clasificación. En el grupo 0, grupo que representa los fracasos, para 1,044 estudiantes que fracasaron, el modelo predijo erróneamente que tendrían éxito, lo que representó un 15.1% de error de mala clasificación para este tipo de situación. En el grupo 1, grupo que representa el éxito, para 1500 que tuvieron éxito, el modelo predijo erróneamente que fracasarían y los clasificó en el grupo 0, lo que representó un 21.7 % de error de mala clasificación por esta otra situación. Por lo tanto la tasa total de mala clasificación fue de $2544/6924 = 36.7\%$, ligeramente mayor que la obtenida al aplicar el Análisis Discriminante Lineal.

Un falso positivo ocurre cuando se predice éxito del estudiante y en realidad el estudiante fracasa. Nuestra regla de decisión predijo que 3,750 de los estudiantes tendría éxito. Esta predicción fue errónea para 1,044 de los estudiantes, por lo que la tasa del falso positivo es $1044/3750 = 27.8\%$.

Un falso negativo ocurre cuando se predice el fracaso del estudiante y en realidad el estudiante obtiene éxito. Nuestra regla de decisión predijo que 3,174 de los estudiantes

fracasarían en su primer curso de matemáticas. Esta predicción fue errónea para 1,500 de los estudiantes, por lo que la tasa del falso negativo es $1500/3174 = 47.3\%$.

Este falso negativo no debe ser motivo de preocupación porque cuando se utilice el modelo para predecir, estos estudiantes al ser clasificados como fracaso se les proveería ayuda adicional, o se les advertiría del riesgo de fracasar que tienen de acuerdo a sus antecedentes académicos por lo que con su esfuerzo podría tener éxito.

La sensitividad es la proporción de verdaderos éxitos identificados por el modelo del total de estudiantes que lograron el éxito en su primer curso de matemáticas.

Sensitividad = $S = 2706/4206 = 64.3\%$.

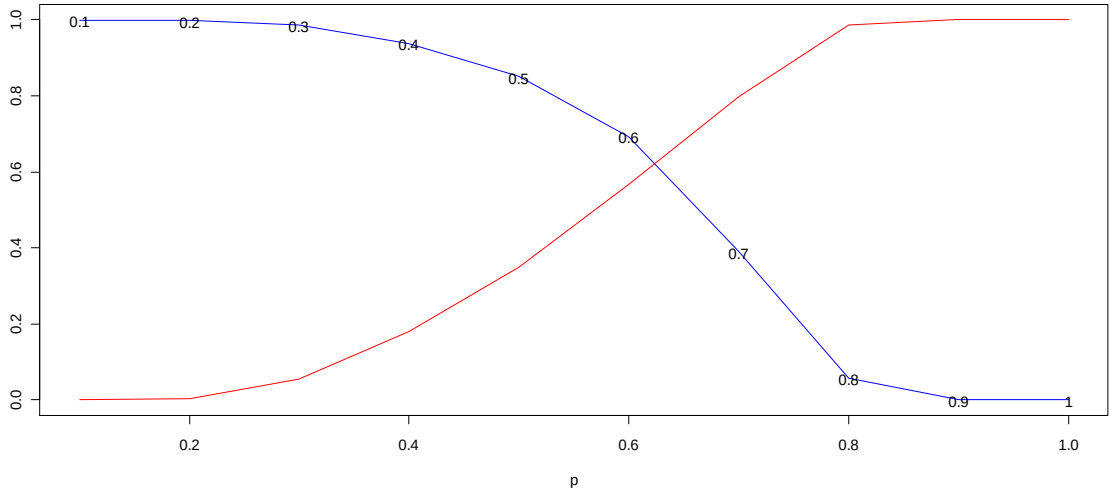
La especificidad es la proporción de verdaderos fracasos identificados por el modelo del total de fracasos.

Especificidad = $E = 1674/2718 = 61.6\%$.

La Figura 4.3 nos presenta la grafica de sensitividad y especificidad buscamos que ambas estén en su punto optimo llamado punto de corte, esto es maximizar la sensitividad al mismo tiempo que maximizamos la especificidad. El punto de corte es el punto donde ambas curvas concurren, esta gráfica nos presenta a 0.62 como punto de corte.

Análisis de Sensitividad y Especificidad

FIGURA 4.3
Gráfica de Sensitividad y Especificidad



El valor de corte para este modelo de regresión logística es 0.62, según calculado en SPSS. El valor de corte es el que nos indica en qué grupo será clasificado el estudiante. Es decir, si al sustituir los datos de un estudiante en la función de regresión logística el resultado en un valor mayor que 0.62, el estudiante será clasificado en el grupo 1, el de éxito. Y si al aplicar la función a los datos de un estudiante el valor resultante es menor que 0.62, el estudiante será clasificado en el grupo 0, el de fracaso.

Variables significativas en el modelo

Hipótesis:

$H_0 =$ la variable no es significativa para el modelo de regresión

$H_1 =$ la variable es significativa para el modelo de regresión

TABLA 4.13
Valor p para cada variable

Variable		Valor p
x_1	PROMEDIO_ESC_SUP	**.000**
x_2	VERBAL_APT	.127
x_3	MATE_APT	**.000**
x_4	INGL_APR	.124
x_5	MATE_APR	**.008**
x_6	ESPA_APR	.083

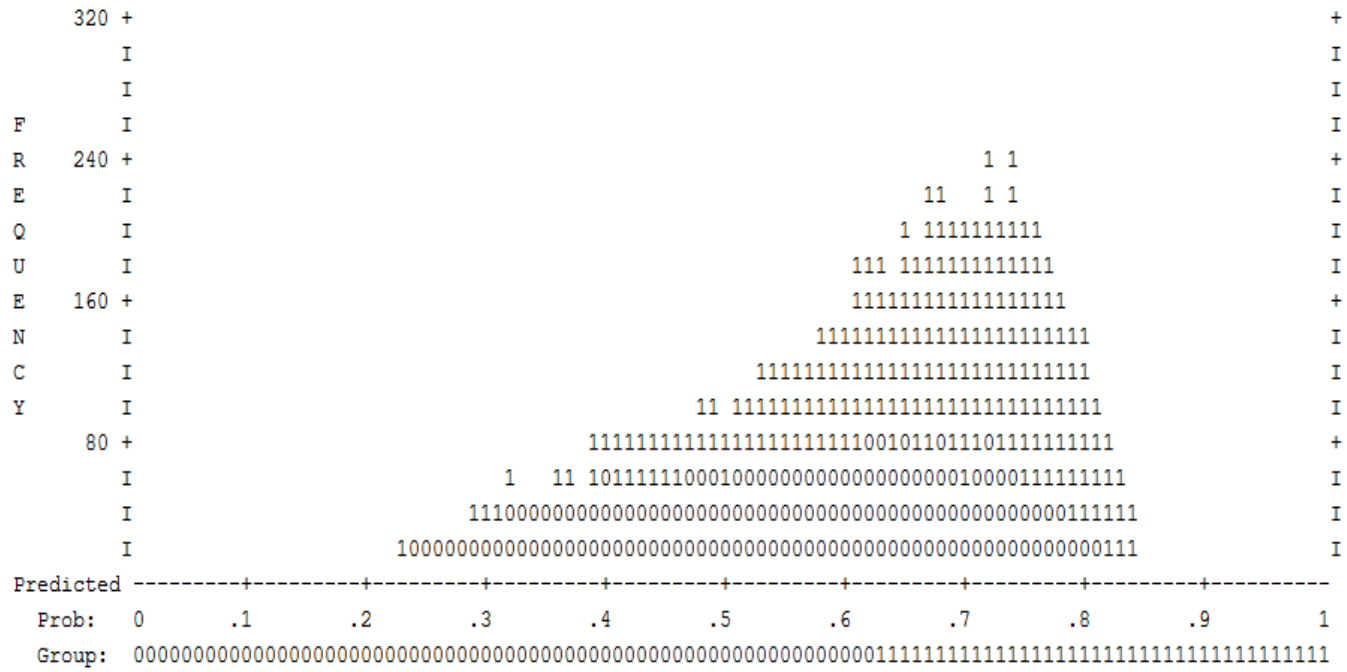
La tabla 4.13 nos muestra los valores de p para cada una de las variables del modelo, notamos que el promedio de escuela superior y las puntuaciones de aptitud matemática y aprovechamiento matemático son las únicas variables significativas en el modelo.

La tabla 4.14 muestra un dato (caso 2153) el cual utilizando el modelo de regresión logística están mal clasificados, notamos que como el corte el 0.62, si el resultado al evaluar el modelo es menor que 0.62 el dato o estudiante será un posible fracaso, y si al evaluar el modelo obtenemos un valor mayor que 0.62 el estudiante será un posible éxito. Al evaluar la función en los valores de las diferentes variables del caso 2153 obtenemos un valor de $0.024 < 0.62$ por lo que el estudiante es un posible fracaso. La realidad es que el estudiante número 2153 perteneció al grupo de éxito en la primera clase de matemáticas.

Tabla 4.14
Ejemplo de un dato mal clasificado

Caso	Grupo real	Asignado por el modelo
2153	1 = Éxito	0.024 = Fracaso

FIGURA 4.4
Grupo de datos observados y sus respectivas probabilidades predictivas



La Figura 4.4 nos muestra el grupo de datos observados o sea el de las notas reales y sus respectivas probabilidades predictivas. Ya se estableció que probabilidades menores que 0.62 serán considerados fracasos y probabilidades mayores o iguales que 0.62 son consideradas éxito en el primer curso de matemáticas.

4.4 Modelo de regresión logística para variables numéricas significativas

En esta sección se utilizará la técnica multivariada de regresión logística para la obtención un modelo final para el cálculo del éxito o el fracaso de un estudiante en su primer curso de matemáticas. En este modelo la variable de respuesta es la variable categórica que toma valor

0 si el estudiante fracasa, es decir, obtiene una nota de D, F, NP o W en su primer curso de matemáticas o un valor de 1 si el estudiante tiene éxito, es decir, obtiene una nota de A, B, C o P, en su primer curso de matemáticas. En este modelo se utilizan las variables independientes que fueron significativas para el modelo de regresión logística anterior, a este modelo se le conoce como modelo de regresión reducido.

El modelo matemático es el siguiente:

Éxito_Fracaso = f(Promedio de Escuela, Mate Apt, Mate Apr)

Descripción de cada Variable:

Éxito o fracaso: Fracaso = 0, Éxito = 1

Promedio de Escuela: Promedio de escuela superior

Mate Apt: Puntuación de Aptitud matemática en la Prueba de Admisión Universitaria

Mate Apr: Puntuación de Aprovechamiento académico en matemáticas en la Prueba de Admisión Universitaria

La tabla 4.15 muestra los coeficientes de la función de regresión logística para cada una de las variables independientes en el modelo.

TABLA 4.15
Coeficientes para la función de regresión logística reducida

x_1	PROMEDIO_ESC_SUP	1.313
x_2	MATE_APT	.001
x_3	MATE_APR	.003
	Constante	-6.879

Por lo tanto, la función estimada de regresión logística para el éxito o fracaso de los estudiantes es la siguiente:

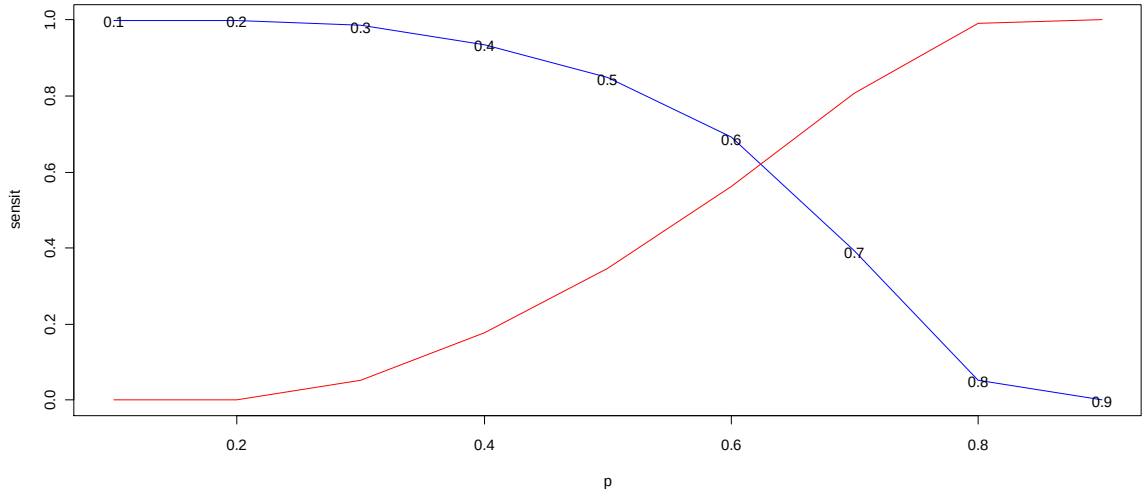
$$p_i = \frac{e^{-6.879+1.313x_1+0.001x_2+0.003x_3}}{1+e^{-6.879+1.313x_1+0.001x_2+0.003x_3}} \qquad (14)$$

donde p_i es la probabilidad de que un estudiante obtenga un valor 0 ó un valor 1 al sustituir sus datos correspondientes a las variables independientes en la ecuación anterior.

La Figura 4.5 nos presenta la grafica de sensitividad y especificidad buscamos que ambas estén en su punto óptimo llamado punto de corte, esto es maximizar la sensitividad al mismo tiempo que maximizamos la especificidad. El punto de corte es el punto donde ambas concurren, esta gráfica nos presenta a 0.65 como punto de corte.

Análisis de Sensitividad y Especificidad

FIGURA 4.5
Gráfica de Sensitividad y Especificidad para el Modelo Reducido



El valor de corte para este modelo de regresión logística es 0.65, según calculado en SPSS. El valor de corte es el que nos indica en qué grupo será clasificado el estudiante. Es decir, si al sustituir los datos de un estudiante en la función de regresión logística el resultado en un valor mayor que 0.65, el estudiante será clasificado en el grupo 1, el de éxito. Y si al aplicar la función a los datos de un estudiante el valor resultante es menor que 0.65, el estudiante será clasificado en el grupo 0, el de fracaso.

Reglas de decisión para clasificar a los estudiantes de nuevo ingreso en cuanto a su posibilidad de tener éxito o fracaso en su primer curso de matemáticas en el RUM:

Si $\frac{e^{-6.879+1.313x_1+0.001x_2+0.003x_3}}{1+e^{-6.879+1.313x_1+0.001x_2+0.003x_3}} \geq 0.65$, entonces se considera que el estudiante tendrá éxito; y si $\frac{e^{-6.879+1.313x_1+0.001x_2+0.003x_3}}{1+e^{-6.879+1.313x_1+0.001x_2+0.003x_3}} < 0.65$, entonces se considera que el estudiante fracasará.

Para clasificar cada estudiante, cada dato, en el éxito o fracaso en su primer curso de matemáticas utilizamos las desigualdades antes mencionadas. Si al evaluar cada variable de un caso se obtiene un número menor que 0.65 esta será clasificada como fracaso. Por el contrario si evaluamos y obtenemos un número mayor o igual que 0.65 entonces este estudiante es clasificado como un éxito en el primer curso de matemáticas.

TABLA 4.16
Clasificación de resultados por error aparente

	Asignados por el modelo			
Datos Reales		0 Fracaso	1 Éxito	Totales
	0 Fracaso	1866	852	2718 (39.3 %)
	1 Éxito	1850	2356	4206 (60.7 %)
	Totales	3174 (53.7 %)	3208 (46.3 %)	6924 (100 %)

La tabla 4.16 muestra el número de datos en cada clasificación. En el grupo 0, grupo que representa los fracasos, para 852 estudiantes que fracasaron, el modelo predijo erróneamente que tendrían éxito, lo que representó un 12.3 % de error de mala clasificación para este tipo de situación. En el grupo 1, grupo que representa el éxito, para 1850 que tuvieron éxito, el modelo predijo erróneamente que fracasarían y los clasificó en el grupo 0, lo que representó un 26.7 % de error de mala clasificación por esta otra situación. Por lo tanto la tasa total de mala clasificación fue de $2702/6924 = 39.0\%$, ligeramente mayor que la obtenida al aplicar el Análisis Discriminante Lineal.

Un falso positivo ocurre cuando se predice éxito del estudiante y en realidad el estudiante fracasa. Nuestra regla de decisión predijo que 3,208 de los estudiantes tendría éxito. Esta predicción fue errónea para 852 de los estudiantes, por lo que la tasa del falso positivo es $852/3208 = 26.6\%$.

Un falso negativo ocurre cuando se predice el fracaso del estudiante y en realidad el estudiante obtiene éxito. Nuestra regla de decisión predijo que 3,716 de los estudiantes

fracasarían en su primer curso de matemáticas. Esta predicción fue errónea para 1,850 de los estudiantes, por lo que la tasa del falso negativo es $1850/3716 = 49.8\%$.

Este falso negativo no debe ser motivo de preocupación porque cuando se utilice el modelo para predecir, estos estudiantes al ser clasificados como fracaso se les proveería ayuda adicional, o se les advertiría del riesgo de fracasar que tienen de acuerdo a sus antecedentes académicos por lo que con su esfuerzo podría tener éxito.

La sensitividad es la proporción de verdaderos éxitos identificados por el modelo del total de estudiantes que lograron el éxito en su primer curso de matemáticas.

Sensitividad = $S = 2356/4206 = 56.0\%$.

La especificidad es la proporción de verdaderos fracasos identificados por el modelo del total de fracasos.

Especificidad = $E = 1866/2718 = 68.7\%$.

4.5 Regresión Logística para variables numéricas y categóricas

En esta sección se utilizará la técnica multivariada de regresión logística para variables las mismas variables numéricas del modelo antes estudiado, añadiendo las siguientes variables categóricas: Tipo de escuela (pública o privada) y Código de escuela de procedencia.

En esta sección se utilizará la técnica multivariada de regresión logística para un modelo final para el cálculo del éxito o el fracaso de un estudiante en su primer curso de matemáticas. La variable de respuesta es la variable dicotómica que toma valor 0 si el estudiante fracasa, es

decir, obtiene D, F, NP o W, en su primer curso de matemáticas, o 1 si el estudiante tiene éxito, es decir, obtiene A, B, C o P, en su primer curso de matemáticas.

El modelo matemático es el siguiente:

Éxito_Fracaso = f(Promedio de Escuela, Verbal Apt, Mate Apt, Ingles Apr, Mate Apr, Espa Apr, tipo de escuela, código de la escuela de procedencia)

Descripción de cada Variable:

Éxito o fracaso: Fracaso = 0, Éxito = 1

Promedio de Escuela: Promedio de escuela superior

Verbal Apt: Puntuación de Aptitud verbal en la Prueba de Admisión Universitaria

Mate Apt: Puntuación de Aptitud matemática en la Prueba de Admisión Universitaria

Ingles Apr: Puntuación de Aprovechamiento académico en inglés en la Prueba de Admisión Universitaria

Mate Apr: Puntuación de Aprovechamiento académico en matemáticas en la Prueba de Admisión Universitaria

Espa Apr: Puntuación de Aprovechamiento académico en español en la Prueba de Admisión Universitaria.

Tipo de Escuela: Privada = 1, Pública = 0; Variable Categórica

Código escuela Superior: Código de la escuela de procedencia; Variable Categórica

La tabla 4.17 representa los coeficientes de la función de Regresión Logística.

TABLA 4.17
Coeficientes para la función logística

x_1	PROMEDIO_ESC_SUP	1.475
x_2	TIPO_DE_ESCUELA(1)	.202
x_3	VERBAL_APT	-.001
x_4	MATE_APT	.003
x_5	INGL_APR	-.001
x_6	MATE_APR	.001
x_7	ESPA_APR	.001
x_8	CODIF_ESCUELA_SUPERIOR	.000
	Constante	-7.835

La función logística para el éxito o fracaso de los estudiantes es la siguiente:

$$p_i = \frac{e^{-7.835+1.475x_1+0.202x_2-0.001x_3+0.003x_4-0.001x_5+0.001x_6+0.001x_7}}{1+e^{-7.835+1.475x_1+0.202x_2-0.001x_3+0.003x_4-0.001x_5+0.001x_6+0.001x_7}} \qquad (15)$$

Reglas de decisión para clasificar a los estudiantes de nuevo ingreso en cuanto a su posibilidad de tener éxito o fracaso en su primer curso de matemáticas en el RUM:

Si $\frac{e^{-7.835+1.475x_1+0.202x_2-0.001x_3+0.003x_4-0.001x_5+0.001x_6+0.001x_7}}{1+e^{-7.835+1.475x_1+0.202x_2-0.001x_3+0.003x_4-0.001x_5+0.001x_6+0.001x_7}} \geq 0.62$, entonces se considera que el estudiante tendrá éxito; y si $\frac{e^{-7.835+1.475x_1+0.202x_2-0.001x_3+0.003x_4-0.001x_5+0.001x_6+0.001x_7}}{1+e^{-7.835+1.475x_1+0.202x_2-0.001x_3+0.003x_4-0.001x_5+0.001x_6+0.001x_7}} < 0.62$, entonces se considera que el estudiante fracasará.

Para clasificar cada estudiante, cada dato, en el éxito o fracaso en su primer curso de matemáticas utilizamos las desigualdades antes mencionadas. Si al evaluar cada variable de un caso se obtiene un número menor que 0.62 esta será clasificada como fracaso. Por el contrario si evaluamos y obtenemos un número mayor o igual que 0.62 entonces este estudiante es clasificado como un éxito en el primer curso de matemáticas.

TABLA 4.18
Clasificación de resultados por error aparente

	Asignados por el modelo			
Datos Reales		0 Fracaso	1 Éxito	Totales
	0 Fracaso	1703	1015	2718 (39.3 %)
	1 Éxito	1520	2686	4206 (60.7 %)
	Totales	3223 (46.6 %)	3701 (53.4 %)	6924 (100 %)

La tabla 4.18 muestra el número de datos que fueron mal clasificados, en el grupo 0, grupo que representa los fracasos, 1015 estudiantes fueron asignados incorrectamente por el modelo al grupo 1, el de los éxitos. En el grupo 1, grupo que representa el éxito, 1520 estudiantes fueron asignados incorrectamente por el modelo al grupo 0, el grupo de los fracasos. Por lo tanto la tasa total de mala clasificación fue de $2535/6924 = 36.6\%$, ligeramente mayor que la obtenida al aplicar el Análisis Discriminante Lineal.

Un falso positivo ocurre cuando se predice éxito del estudiante y en realidad el estudiante fracasa. Nuestra regla de decisión predijo que 3,701 de los estudiantes tendría éxito. Esta predicción fue errónea para 1,015 de los estudiantes, por lo que la tasa del falso positivo es $1015/3701 = 27.4\%$.

Un falso negativo ocurre cuando se predice el fracaso del estudiante y en realidad el estudiante obtiene éxito. Nuestra regla de decisión predijo que 3,223 de los estudiantes fracasarían en su primer curso de matemáticas. Esta predicción fue errónea para 1,520 de los estudiantes, por lo que la tasa del falso negativo es $1520/3223 = 47.2\%$.

Este falso negativo no debe ser motivo de preocupación porque cuando se utilice el modelo para predecir, estos estudiantes al ser clasificados como fracaso se les proveería ayuda adicional, o se les advertiría del riesgo de fracasar que tienen de acuerdo a sus antecedentes académicos por lo que con su esfuerzo podría tener éxito.

La sensitividad es la proporción de verdaderos éxitos identificados por el modelo del total de estudiantes que lograron el éxito en su primer curso de matemáticas.

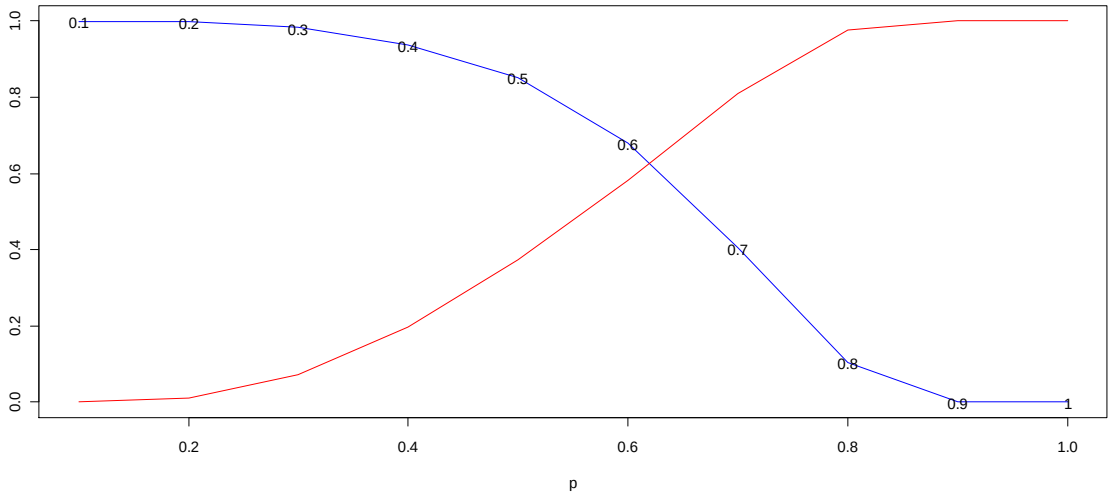
Sensitividad = $S = 2686/4206 = 63.9\%$.

La especificidad es la proporción de verdaderos fracasos identificados por el modelo del total de fracasos.

Especificidad = $E = 1703/2718 = 62.7\%$.

Análisis de Sensitividad y Especificidad

FIGURA 4.6
Gráfica de Sensitividad y Especificidad



El valor de corte para este modelo de regresión logística es 0.62, según calculado en SPSS. El valor de corte es el que nos indica en qué grupo será clasificado el estudiante, según sus datos correspondientes a las variables independientes. Es decir, si al aplicar la función de regresión logística a los datos de un estudiante particular el valor resultante es mayor o igual que 0.62, el estudiante será clasificado en el grupo 1, el de los éxitos. Y si al evaluar la función el valor resultante es menor que 0.62 el estudiante será clasificado en el grupo 0, el de los fracasos.

Variables significativas en el modelo

Hipotesis:

$H_0 =$ la variable no es significativa para el modelo de regresión

$H_1 =$ la variable es significativa para el modelo de regresión

TABLA 4.19
Valor p para cada variable

	Variable	Valor p
x_1	PROMEDIO_ESC_SUP	**.000**
x_2	TIPO_DE_ESCUELA(1)	.241
x_3	VERBAL_APT	.078
x_4	MATE_APT	**.000**
x_5	INGL_APR	**.001**
x_6	MATE_APR	**.015**
x_7	ESPA_APR	.151
x_8	CODIF_ESCUELA_SUPERIOR	**.000**

La tabla 4.19 nos muestra los valores de p para cada una de las variables del modelo, notamos que el promedio de escuela superior, las puntuaciones de aptitud matemática, aprovechamiento en matemáticas y en ingles y codificación de la escuela de procedencia son las únicas variables significativas en el modelo.

Tabla 4.20
Lista de un dato mal clasificado

Caso	Grupo real	Asignado por el modelo
113	0 = Fracaso	0.883 = Éxito
4871	1 = Éxito	0.014 = Fracaso
6845	1 = Éxito	0.097 = Fracaso

La tabla 4.20 muestra cuatro datos (los casos 113, 4871, 6845) los cuales utilizando el modelo de regresión logística están mal clasificados, notamos que como el corte el 0.62 si el

resultado al evaluar la función el resultado es menor que 0.62 el modelo predice que el estudiante será un posible fracaso, y si al evaluar la función el resultado es un valor mayor que 0.62 el estudiante será un posible éxito.

4.6 Regresión Logística para variables numéricas y categóricas significativas

En esta sección se utilizará la técnica multivariada de regresión logística para variables numéricas y categóricas que en el modelo anterior resultaron significativas.

El modelo matemático es el siguiente:

Éxito_Fracaso = f(Promedio de Escuela, Mate Apt, Ingles Apr, Mate Apr, Código de la escuela de procedencia)

Descripción de cada Variable:

Éxito o fracaso: Fracaso = 0, Éxito = 1

Promedio de Escuela: Promedio de escuela superior

Mate Apt: Puntuación de Aptitud matemática en la Prueba de Admisión Universitaria

Ingles Apr: Puntuación de Aprovechamiento académico en inglés en la Prueba de Admisión Universitaria

Mate Apr: Puntuación de Aprovechamiento académico en matemáticas en la Prueba de Admisión Universitaria

Código escuela Superior: Código de la escuela de procedencia; Variable Categórica

La tabla 4.21 representa los coeficientes de la función de Regresión Logística.

TABLA 4.21
Coeficientes para la función logística

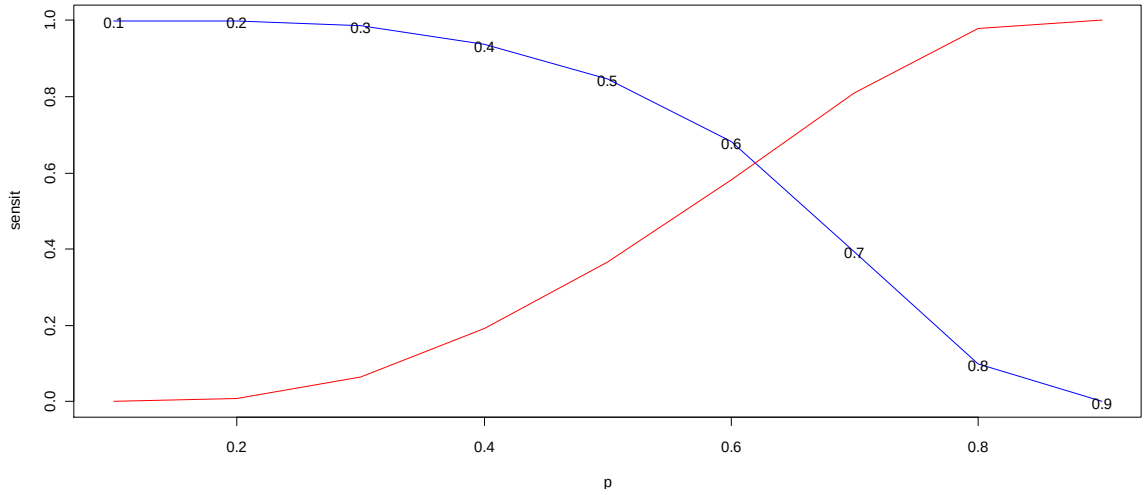
x_1	PROMEDIO_ESC_SUP	1.479
x_2	MATE_APT	.003
x_3	INGL_APR	-.001
x_4	MATE_APR	.001
x_5	CODIF_ESCUELA_SUPERIOR	.000
	Constante	-7.516

La función logística para el éxito o fracaso de los estudiantes es la siguiente:

$$p_i = \frac{e^{-7.516+1.479x_1+0.003x_2-0.001x_3+0.001x_4}}{1+e^{-7.516+1.479x_1+0.003x_2-0.001x_3+0.001x_4}} \qquad (16)$$

La Figura 4.7 nos presenta la grafica de sensitividad y especificidad buscamos que ambas estén en su punto optimo llamado punto de corte, esto es maximizar la sensitividad al mismo tiempo que maximizamos la especificidad. El punto de corte es el punto donde ambas concurren, esta grafica nos presenta a 0.63 como punto de corte.

FIGURA 4.7
Gráfica de Sensitividad y Especificidad



El valor de corte para este modelo de regresión logística es 0.63, según calculado en SPSS. El valor de corte es el que nos indica en qué grupo será clasificado el estudiante. Es decir, si al sustituir los datos de un estudiante en la función de regresión logística el resultado en un valor mayor que 0.63, el estudiante será clasificado en el grupo 1, el de éxito. Y si al aplicar la función a los datos de un estudiante el valor resultante es menor que 0.63, el estudiante será clasificado en el grupo 0, el de fracaso.

Reglas de decisión para clasificar a los estudiantes de nuevo ingreso en cuanto a su posibilidad de tener éxito o fracaso en su primer curso de matemáticas en el RUM:

Si $\frac{e^{-7.516+1.479x_1+0.003x_2-0.001x_3+0.001x_4}}{1+e^{-7.516+1.479x_1+0.003x_2-0.001x_3+0.001x_4}} \geq 0.63$, entonces se considera que el estudiante tendrá éxito; y si $\frac{e^{-7.516+1.479x_1+0.003x_2-0.001x_3+0.001x_4}}{1+e^{-7.516+1.479x_1+0.003x_2-0.001x_3+0.001x_4}} < 0.63$, entonces se considera que el estudiante fracasará.

Para clasificar cada estudiante, cada dato, en el éxito o fracaso en su primer curso de matemáticas utilizamos las desigualdades antes mencionadas. Si al evaluar cada variable de un caso se obtiene un número menor que 0.63 esta será clasificada como fracaso. Por el contrario si evaluamos y obtenemos un número mayor o igual que 0.63 entonces este estudiante es clasificado como un éxito en el primer curso de matemáticas.

TABLA 4.22
Clasificación de resultados por error aparente

	Asignados por el modelo			
Datos Reales		0 Fracaso	1 Éxito	Totales
	0 Fracaso	1760	958	2718 (39.3 %)
	1 Éxito	1628	2578	4206 (60.7 %)
	Totales	3388 (48.9 %)	3536 (51.1 %)	6924 (100 %)

La tabla 4.22 muestra el número de datos en cada clasificación. En el grupo 0, grupo que representa los fracasos, para 958 estudiantes que fracasaron, el modelo predijo erróneamente que tendrían éxito, lo que representó un 13.8 % de error de mala clasificación para este tipo de situación. En el grupo 1, grupo que representa el éxito, para 1,628 que tuvieron éxito, el modelo predijo erróneamente que fracasarían y los clasificó en el grupo 0, lo que representó un 23.5 % de error de mala clasificación por esta otra situación. Por lo tanto la tasa total de mala clasificación fue de $2586/6924 = 37.3\%$.

Un falso positivo ocurre cuando se predice éxito del estudiante y en realidad el estudiante fracasa. Nuestra regla de decisión predijo que 3,536 de los estudiantes tendría éxito. Esta predicción fue errónea para 958 de los estudiantes, por lo que la tasa del falso positivo es $958/3536 = 27.1\%$.

Un falso negativo ocurre cuando se predice el fracaso del estudiante y en realidad el estudiante obtiene éxito. Nuestra regla de decisión predijo que 3,388 de los estudiantes

fracasarían en su primer curso de matemáticas. Esta predicción fue errónea para 1,628 de los estudiantes, por lo que la tasa del falso negativo es $1628/3388 = 51.9\%$.

Este falso negativo no debe ser motivo de preocupación porque cuando se utilice el modelo para predecir, estos estudiantes al ser clasificados como fracaso se les proveería ayuda adicional, o se les advertiría del riesgo de fracasar que tienen de acuerdo a sus antecedentes académicos por lo que con su esfuerzo podría tener éxito.

La sensitividad es la proporción de verdaderos éxitos identificados por el modelo del total de estudiantes que lograron el éxito en su primer curso de matemáticas.

Sensitividad = $S = 2578/4206 = 61.3\%$.

La especificidad es la proporción de verdaderos fracasos identificados por el modelo del total de fracasos.

Especificidad = $E = 1760/2718 = 64.8\%$.

4.7 Regresión Logística para variables numéricas para estudiantes de cursos avanzados

En esta sección se utilizará la técnica multivariada de regresión logística las mismas variables numéricas del modelo antes estudiado, la diferencia es que en este estudio se tomaron en cuenta solo los estudiantes que tomaron cursos como: MATE 3005, MATE 3018, MATE 3021, MATE 3031, MATE 3143, MATE 3151, MATE3171 y MATE 3172. Todos estos cursos tienen en común contenidos de Precálculo y cálculo, algunos de ellos con diferentes enfoques.

En esta sección se utilizará la técnica multivariada de regresión logística para determinar un modelo final para el cálculo del éxito o el fracaso de un estudiante en su primer curso de matemáticas. La variable de respuesta es la variable dicotómica que toma valor 0 si el estudiante fracasa, es decir, obtiene D, F, NP o W, en su primer curso de matemáticas, o 1 si el estudiante tiene éxito, es decir, obtiene A, B, C o P, en su primer curso de matemáticas.

TABLA 4.23
Proporción de Éxito y fracaso de estudiantes

	Numero de estudiantes	Proporción
Éxito	2510	60.9 %
Fracaso	1609	39.1 %
Total	4119	100 %

El modelo matemático es el siguiente:

Éxito_Fracaso = f(Promedio de Escuela, Verbal Apt, Mate Apt, Ingles Apr, Mate Apr, Espa Apr)

Descripción de cada Variable:

Éxito o fracaso: Fracaso = 0, Éxito = 1

Promedio de Escuela: Promedio de escuela superior

Verbal Apt: Puntuación de Aptitud verbal en la Prueba de Admisión Universitaria

Mate Apt: Puntuación de Aptitud matemática en la Prueba de Admisión Universitaria

Ingles Apr: Puntuación de Aprovechamiento académico en inglés en la Prueba de Admisión Universitaria

Mate Apr: Puntuación de Aprovechamiento académico en matemáticas en la Prueba de Admisión Universitaria

Espa Apr: Puntuación de Aprovechamiento académico en español en la Prueba de Admisión Universitaria.

La tabla 4.24 representa los coeficientes de la función de Regresión Logística.

TABLA 4.24
Coeficientes para la función logística

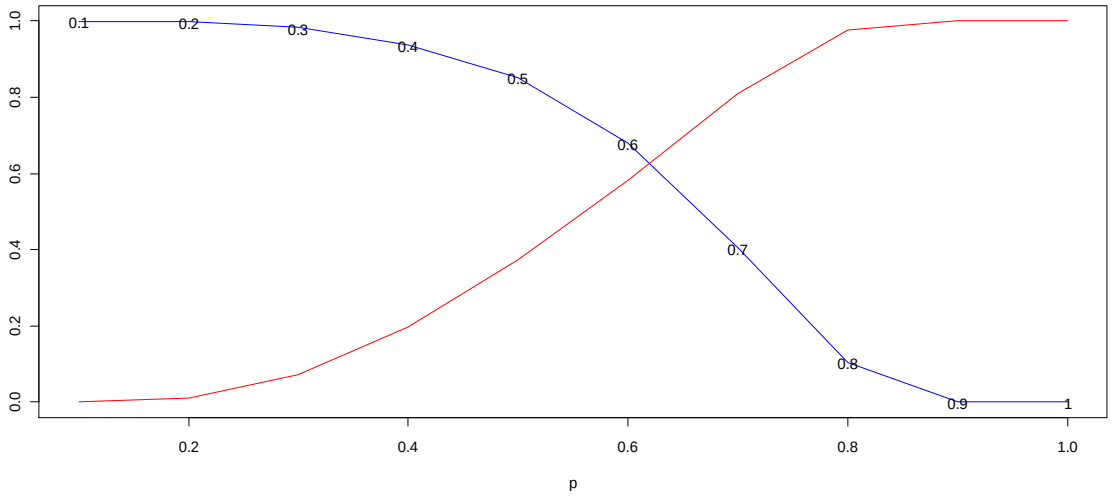
x_1	PROMEDIO_ESC_SUP	1.536
x_2	VERBAL_APT	.000
x_3	MATE_APT	.004
x_4	INGL_APR	.000
x_5	MATE_APR	.003
x_6	ESPA_APR	.001
	Constante	-10.460

La función logística para el éxito o fracaso de los estudiantes es la siguiente:

$$p_i = \frac{e^{-10.460+1.536x_1+0.004x_3+0.003x_5+0.001x_6}}{1+e^{-10.460+1.536x_1+0.004x_3+0.003x_5+0.001x_6}} \qquad (17)$$

Análisis de Sensitividad y Especificidad

FIGURA 4.8
Gráfica de Sensitividad y Especificidad



El valor de corte para este modelo de regresión logística es 0.63, según calculado en SPSS. El valor de corte es el que nos indica en qué grupo será clasificado el estudiante, según sus datos correspondientes a las variables independientes. Es decir, si al aplicar la función de regresión logística a los datos de un estudiante particular el valor resultante es mayor o igual que 0.63, el estudiante será clasificado en el grupo 1, el de los éxitos. Y si al evaluar la función el valor resultante es menor que 0.63 el estudiante será clasificado en el grupo 0, el de los fracasos.

Reglas de decisión para clasificar a los estudiantes de nuevo ingreso en cuanto a su posibilidad de tener éxito o fracaso en su primer curso de matemáticas en el RUM:

Si $\frac{e^{-10.460+1.536x_1+0.004x_3+0.003x_5+0.001x_6}}{1+e^{-10.460+1.536x_1+0.004x_3+0.003x_5+0.001x_6}} \geq 0.63$, entonces se considera que el estudiante tendrá éxito; y si $\frac{e^{-10.460+1.536x_1+0.004x_3+0.003x_5+0.001x_6}}{1+e^{-10.460+1.536x_1+0.004x_3+0.003x_5+0.001x_6}} > 0.63$, entonces se considera que el estudiante fracasará.

Para clasificar cada estudiante, cada dato, en el éxito o fracaso en su primer curso de matemáticas utilizamos las desigualdades antes mencionadas. Si al evaluar cada variable de un caso se obtiene un número menor que 0.63 esta será clasificada como fracaso. Por el contrario si evaluamos y obtenemos un número mayor o igual que 0.63 entonces este estudiante es clasificado como un éxito en el primer curso de matemáticas.

TABLA 4.25
Clasificación de resultados por error aparente

	Asignados por el modelo			
Datos Reales		0 Fracaso	1 Éxito	Totales
	0 Fracaso	1053	556	1609 (39.1%)
	1 Éxito	921	1589	2510 (60.9 %)
	Totales	1974 (47.9 %)	2145 (52.1 %)	4119 (100 %)

La tabla 4.25 muestra el número de datos que fueron mal clasificados, en el grupo 0, grupo que representa los fracasos, 556 estudiantes fueron asignados incorrectamente por el modelo al grupo 1, grupo de los éxitos. En el grupo 1, grupo que representa el éxito, 921 estudiantes fueron asignados incorrectamente por el modelo al grupo 0, el grupo de los fracasos. Por lo tanto la tasa total de mala clasificación fue de $1477/4119 = 35.6\%$.

Un falso positivo ocurre cuando se predice éxito del estudiante y en realidad el estudiante fracasa. Nuestra regla de decisión predijo que 2,145 de los estudiantes tendrían éxito. Esta predicción fue errónea para 556 de los estudiantes, por lo que la tasa del falso positivo es $556/2145 = 25.9\%$.

Un falso negativo ocurre cuando se predice el fracaso del estudiante y en realidad el estudiante obtiene éxito. Nuestra regla de decisión predijo que 1,974 de los estudiantes fracasarían en su primer curso de matemáticas. Esta predicción fue errónea para 921 de los estudiantes, por lo que la tasa del falso negativo es $921/1974 = 46.7\%$.

Este falso negativo no debe ser motivo de preocupación porque cuando se utilice el modelo para predecir, estos estudiantes al ser clasificados como fracaso se les proveería ayuda adicional, o se les advertiría del riesgo de fracasar que tienen de acuerdo a sus antecedentes académicos por lo que con su esfuerzo podría tener éxito.

La sensitividad es la proporción de verdaderos éxitos identificados por el modelo del total de estudiantes que lograron el éxito en su primer curso de matemáticas.

Sensitividad = $S = {}^{1589}\!/_{2510} = 63.3\%$.

La especificidad es la proporción de verdaderos fracasos identificados por el modelo del total de fracasos.

Especificidad = $E = {}^{1053}\!/_{1609} = 65.4\%$.

Variables significativas en el modelo

Hipótesis:

$H_0 =$ la variable no es significativa para el modelo de regresión

$H_1 =$ la variable es significativa para el modelo de regresión

TABLA 4.26
Valor p para cada variable

Variable		Valor p
x_1	PROMEDIO_ESC_SUP	.000
x_2	VERBAL_APT	.759
x_3	MATE_APT	.000
x_4	INGL_APR	.588
x_5	MATE_APR	.001
x_6	ESPA_APR	.299

La tabla 4.26 nos muestra los valores de p para cada una de las variables del modelo, notamos que el promedio de escuela superior, las puntuaciones de aptitud matemática y aprovechamiento en matemáticas son las únicas variables significativas en el modelo.

4.8 Regresión Logística para variables numéricas significativas para estudiantes de cursos avanzados

En esta sección se utilizará la técnica multivariada de regresión logística para variables numéricas que en el modelo anterior resultaron significativas.

El modelo matemático es el siguiente:

Éxito_Fracaso = f(Promedio de Escuela, Mate Apt, Mate Apr)

Descripción de cada Variable:

Éxito o fracaso: Fracaso = 0, Éxito = 1

Promedio de Escuela: Promedio de escuela superior

Mate Apt: Puntuación de Aptitud matemática en la Prueba de Admisión Universitaria

Mate Apr: Puntuación de Aprovechamiento académico en matemáticas en la Prueba de Admisión Universitaria

La tabla 4.27 representa los coeficientes de la función de Regresión Logística.

TABLA 4.27
Coeficientes para la función logística

x_1	PROMEDIO_ESC_SUP	1.560
x_2	MATE_APT	.004
x_3	MATE_APR	.003
	Constante	-10.234

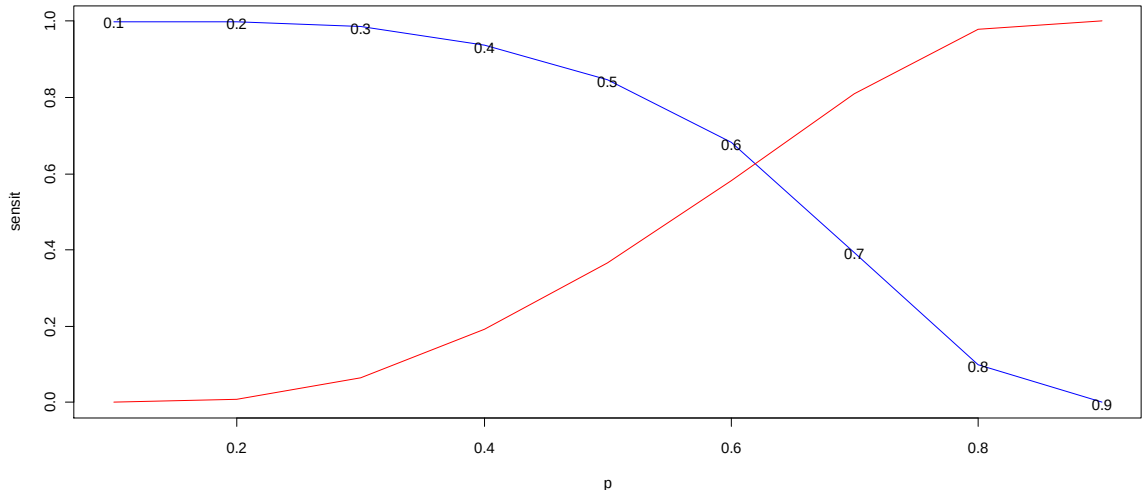
La función logística para el éxito o fracaso de los estudiantes es la siguiente:

$$p_i = \frac{e^{-10.234+1.560x_1+0.004x_2+0.003x_3}}{1+e^{-10.234+1.560x_1+0.004x_2+0.003x_3}} \quad (18)$$

La Figura 4.9 nos presenta la grafica de sensitividad y especificidad buscamos que ambas estén en su punto optimo llamado punto de corte, esto es maximizar la sensitividad al mismo tiempo que maximizamos la especificidad. El punto de corte es el punto donde ambas concurren, esta grafica nos presenta a 0.63 como punto de corte.

Análisis de Sensitividad y Especificidad

FIGURA 4.9
Gráfica de Sensitividad y Especificidad



El valor de corte para este modelo de regresión logística es 0.63, según calculado en SPSS. El valor de corte es el que nos indica en qué grupo será clasificado el estudiante. Es decir, si al sustituir los datos de un estudiante en la función de regresión logística el resultado en un valor mayor que 0.63, el estudiante será clasificado en el grupo 1, el de éxito. Y si al aplicar la función a los datos de un estudiante el valor resultante es menor que 0.63, el estudiante será clasificado en el grupo 0, el de fracaso.

Reglas de decisión para clasificar a los estudiantes de nuevo ingreso en cuanto a su posibilidad de tener éxito o fracaso en su primer curso de matemáticas en el RUM:

Si $\frac{e^{-10.234+1.560x_1+0.004x_2+0.003x_3}}{1+e^{-10.234+1.560x_1+0.004x_2+0.003x_3}} \geq 0.63$, entonces se considera que el estudiante tendrá éxito; y si $\frac{e^{-10.234+1.560x_1+0.004x_2+0.003x_3}}{1+e^{-10.234+1.560x_1+0.004x_2+0.003x_3}} < 0.63$, entonces se considera que el estudiante fracasará.

Para clasificar cada estudiante, cada dato, en el éxito o fracaso en su primer curso de matemáticas utilizamos las desigualdades antes mencionadas. Si al evaluar cada variable de un caso se obtiene un número menor que 0.63 esta será clasificada como fracaso. Por el contrario si evaluamos y obtenemos un número mayor o igual que 0.63 entonces este estudiante es clasificado como un éxito en el primer curso de matemáticas.

TABLA 4.28
Clasificación de resultados por error aparente

	Asignados por el modelo			
Datos Reales		0 Fracaso	1 Éxito	Totales
	0 Fracaso	1048	561	1609 (39.1 %)
	1 Éxito	920	1590	2510 (60.9 %)
	Totales	1968 (47.8 %)	2151 (52.2 %)	4119 (100 %)

La tabla 4.28 muestra el número de datos en cada clasificación. Tenemos entonces que la tasa total de mala clasificación fue de $1481/4119 = 35.9\%$.

Un falso positivo ocurre cuando se predice éxito del estudiante y en realidad el estudiante fracasa. Nuestra regla de decisión predijo que 2,151 de los estudiantes tendría éxito. Esta predicción fue errónea para 561 de los estudiantes, por lo que la tasa del falso positivo es $561/2151 = 26.1\%$.

Un falso negativo ocurre cuando se predice el fracaso del estudiante y en realidad el estudiante obtiene éxito. Nuestra regla de decisión predijo que 1,968 de los estudiantes fracasarían en su primer curso de matemáticas. Esta predicción fue errónea para 920 de los estudiantes, por lo que la tasa del falso negativo es $920/1968 = 46.7\%$.

Este falso negativo no debe ser motivo de preocupación porque cuando se utilice el modelo para predecir, estos estudiantes al ser clasificados como fracaso se les proveería ayuda adicional, o se les advertiría del riesgo de fracasar que tienen de acuerdo a sus antecedentes académicos por lo que con su esfuerzo podría tener éxito.

La sensitividad es la proporción de verdaderos éxitos identificados por el modelo del total de estudiantes que lograron el éxito en su primer curso de matemáticas.

Sensitividad = $S = 1590/2510 = 63.3\%$.

La especificidad es la proporción de verdaderos fracasos identificados por el modelo del total de fracasos.

Especificidad = $E = 1048/1609 = 65.1\%$.

CAPÍTULO 5
CONCLUSIONES

Luego de analizar los hallazgos obtenidos al realizar el presente trabajo de investigación se llegó a las siguientes conclusiones:

Al hacer uso del Análisis Discriminante se obtuvo que el promedio de escuela superior, la puntuación de aprovechamiento matemático y la puntuación de aptitud matemática son las variables que más contribuyen a discriminar entre los grupos éxito o fracaso en el primer curso de matemáticas. Además al utilizar la regresión logística para determinar los diferentes modelos predictivos cabe resaltar que en todos los modelos las variables significativas fueron: el promedio de escuela superior, la puntuación de aprovechamiento matemático y la puntuación de aptitud matemática. Podemos concluir que el promedio de escuela superior, la puntuación de aprovechamiento matemático y la puntuación de aptitud matemática son los factores que más inciden en el rendimiento de los estudiantes en su primer curso de matemáticas.

La implantación de modelos de regresión logística y del análisis de discriminante permitió obtener siete modelos predictivos, estos predicen el éxito o fracaso de los estudiantes en su primer curso de matemáticas. El modelo obtenido por análisis de discriminante sólo consistió de variables numéricas tales como: promedio de escuela superior y las puntuaciones obtenidas por el estudiante en la Prueba de Admisión Universitaria, para determinar el éxito o el fracaso y tiene un error de 36.1 %. Por otro lado obtuvimos seis modelos de regresión logística, uno con las mismas variables métricas del modelo

discrimínate el cual obtuvo un error de 36.7 %, por otro lado el modelo de regresión logística, añadiendo las variables categóricas tipo de escuela y escuela de procedencia, obtuvo un error de 36.6%. Los errores de los modelos de regresión logística están entre 35.6% y 39.0%. Se recomienda la utilización de los modelos reducidos con las variables significativas esto pues su obtiene básicamente la misma conclusión con la utilización de los menos datos provistos. No hay diferencia significativa entre los errores de estos modelos.

Utilizando el análisis por conglomerado podemos decir que para identificar a los estudiantes en riesgo de fracasar o tener éxito no hace falta tener en cuenta el programa de estudio al que el estudiante fue admitido, esto posiblemente nos puede encaminar a poder decir, con mas estudios, que la carga académica y los otros cursos que los estudiantes toman a la vez que su primer curso de matemática no afecta el rendimiento de estos en los mismos.

Utilizando la matriz de correlación entre las variables podemos afirmar que las variables mas correlacionadas son las siguientes: IGS se relaciona con las siguientes variables: promedio de escuela superior, puntuación de aptitud matemática, puntuación de aptitud matemática y con el aprovechamiento matemático. Esto tiene mucho sentido pues para calcular el IGS se toma en cuenta el promedio de escuela superior, la puntuación de aptitud verbal y la puntuación de aptitud matemática. También tenemos que la puntuación en la parte aptitud matemática tiene una alta correlación con la puntuación en el aprovechamiento matemático de los estudiantes.

Además debemos resaltar que el promedio de escuela superior tiene correlaciones muy bajas con todas las variables, excepto con el IGS. Esto podría sugerir que las altas correlaciones

entre el IGS y las variables antes mencionadas se deben más al componente de la PEAU que tiene el IGS que al promedio de escuela superior.

Podemos resaltar que de las matrices de correlación tenemos que en el grupo de fracasos se observa la alta correlación negativa entre las puntuaciones en las secciones de aptitud matemática y aprovechamiento matemático. Sin embargo se observa lo opuesto para el grupo de éxitos una alta correlación positiva entre las secciones de aptitud matemática y aprovechamiento matemático.

Un hallazgo interesante es que entre las medias de los grupos de estudiantes que obtuvieron éxito y del grupo de estudiantes que obtuvieron fracasaron en su primer curso de matemáticas las diferencias más significativas se obtuvieron en: las puntuaciones en Aptitud Matemática y de Aprovechamiento Matemático en la PEAU. Además de que todas las medias de las variables para el grupo que tuvo éxito en su primer curso de matemáticas, son más altos que los del grupo que fracasó en su primer curso de matemáticas.

Un hallazgo relevante es que los estudiantes procedentes de las escuelas privadas obtienen los índices de ingreso universitario y puntuaciones en la Prueba de Admisión Universitaria del College Board más altos, sin embargo obtienen promedios de escuela superior más bajos que los estudiantes procedentes de las escuelas públicas. Esto es indicativo de que los estudiantes de las escuelas privadas vienen con un conocimiento mayor, aunque sus notas obtenidas en la escuela superior sean en promedio más bajas. Además esto es que los estudiantes de escuelas públicas ingresan con promedios de escuela más altos, mientras que

sus puntuaciones en la Prueba de Admisión Universitaria del College Board es más bajo, esto que en promedio estos estudiantes provienen con menor conocimiento.

CAPÍTULO 6
TRABAJOS FUTUROS

Como trabajos futuros planteo los siguientes:

- La verificación de ambos modelos con los estudiantes que ingresaron entre los años 2008 al 2010. Esto ayudará a validar los modelos.
- Aplicar ambos modelos a estudiantes que ingresarán en el 2011 para poder tomar acción con aquéllos que sean identificados como posibles fracasos. Esto conllevará realizar un repaso de destrezas básicas, dependiendo del curso en que el estudiante sea matriculado. Dicho repaso debería realizarse en el verano antes de que comience el curso.
- Recomiendo que no se considere en el estudio los estudiantes que obtuvieron D, pues la frontera entre la C y D, es mínima. Dado que curva en el Departamento de Matemáticas es la siguiente: 100 - 90 % A, 89 – 80 % B, 79 – 65 % C, 64 – 60 % D, 59 – 0 % F.

REFERENCIAS

[1] A. Agresti, Categorical Data Analysis. 2da Ed. John Wiley & Sons. 2002

[2] A. Cameron, *Regression analysis of count data*. Ed. Cambridge UniversityPress, United Kingdom. 1998

[3] A. Cuevas y J. Berrendero, Análisis Discriminante: Prácticas con R. Departamento de Matemáticas Universidad Autónoma de Madrid. 2003

[4] A. Rúa, *Búsqueda de Patrones de Rendimiento Académico Mediante: Técnicas de Análisis Multivariante.* Aplicación a 1º E4. http://150.214.55.100/asepuma/ laspalmas2001/ laspalmas/Doco23.PDF Consultado en noviembre de 2006. 2001

[5] Datos provistos por la Oficina de Investigación Institucional y Planificación (OIIP). 2008

[6] D. Belsley, Kuh, y Welsh, R.,Regression Diagnostics. John Wiley, New York. 1980

[7] D. Hosmer y S. Lemeshow, Applied Logistic Regression, Second Edition. 2000

[8] E. Dallas, Applied Multivariate Methods for Data Analysis, Ed. Duxbury Press, California. 1998.

[9] J. Lattin, J. Carroll y P. Green, *Analyzing Multivariate Data*. Thomson Learning, Canada. 2003

[10] Moral, J., *Predicción del Rendimiento Académico Universitario. Perfiles Educativos* V.28 No. 113. Méjico. 2006

[11] J. Quintana Díaz, *Antecedentes académicos (PEAU) de los estudiantes de primer ingreso (1990-2005)*, Estudio no publicado, RUM, 2007.

[12] J. Segura, J. Solís y V. Segura, Análisis de Regresión Logística para Datos Correlacionados Utilizando Tres Procedimientos del Sistema Estadístico SAS. Revista Científica, FCV- LUZ, Vol. XVI, No 3, 282-287. 2006

[13] J. Villavicencio, *Ciclos económicos de Puerto Rico: Uso de modelos y técnicas estadísticas multivariadas,* Tesis de maestría en estadística, no publicada, RUM, 2010.

[14] L. Kaufman y P. Rousseeuw, Finding groups in data: An Introduction to Cluster Analysis. John Wiley & Sons, New York. 1990

[15] M. Arrieta Illarramendi, *Modelo causal del rendimiento en matemáticas* (11-12 años), Enseñanzas de las ciencias, Vol. 16 (1), pp. 63-71. 1996.

[16] M. González, *A Comparison in Cluster Validation Techniques*. Tesis, 2005

[17] N. Draper y H. Smith, Applied Regression Analysis, Third Edition. John Wiley, New York. 1998

[18] O. Castrillón, *Uso de técnicas multivariadas y modelos estadísticos para el análisis del desempeño académico de los estudiantes de Cálculo I - UPRM*, Tesis de maestría en estadística, no publicada, RUM, 2007.

[19] R. Cervini, *Efecto de la "Oportunidad de aprender" sobre el logro en matemáticas en la educación básica argentina*. Revista Electrónica de Investigación Educativa, Vol. 3 (2), 2001.

[20] V. López Vázquez, J. Quintana Díaz, *Estudio del rendimiento de los estudiantes de matemáticas del Recinto Universitario de Mayagüez*, Estudio no publicado, RUM, 2002.

ANEJOS

```
#include <iostream>
#include <string>
#include <math.h>
using namespace std;

int x; // variable global para el menu.

/*****************************************************************************
****
 * Esta funcion verifica si el estudiante tendra exito o fracaso en su primer curso de
*
 * matematica
*
 *
*
 * input: numero real
*
 * ouput: exito o fracaso
*

*****************************************************************************
****/
string exito_fracaso (double r)
{
	if (r >= 0.62 ) // Si r es mayor o igual que 0.62 el estudiante tendra exito.
	{

		return "\nSaludo estudiante! Por ser un estudiante tan comprometido y
sobresaliente\n"
		   "en las áreas de las matemáticas eres un candidato a tener
éxito en tu \n"
		"primer curso de matemáticas. Solo te recomendamos que repases
los conceptos \n"
		"y destrezas básicas para el curso de matemáticas que estas
próximo a comenzar.\n";
	}
	else // Si r es menor que 0.62 el estudiante no tendra exito.
	{
		return "\nSaludo estudiante! Debes repasar los conceptos y destrezas \n"
			"básicas para los cursos de matemáticas que estas próximo a
comenzar.\n"
			"De lo contrario puede verse afectado tu desempeño en el curso.
\n";
	}
} // fin de  exito_fracaso.
```

```
/*****************************************************************************
 * Programa principal                                                        *
 *****************************************************************************/

int main (){
	double x1,
		x2,
		x3,
		x4,
		x5,
		x6,
		x7,
		 y;


	do {

		cout << "\nEntre: \n"
		"	1  para el analisis discriminate; \n"
		"	2  para la regresion logistica con variable numerica; \n"
		"	3  para la regresion logistica con variable numerica y
categorica;\n"
		"	4  para terminar.\n";
		cin >> x;

		switch (x) {
			case 1: // Calcula la funcion descriminante.
			{


				cout << "Entre los coeficientes para la funcion
logistica:"<< endl;
				cout << "Promedio de Escuela: ";
				cin  >> x1;
				cout << "Verbal Apt: ";
				cin  >> x2;
				cout << "Mate Apt: ";
				cin  >> x3;
				cout << "Ingles Apr: ";
				cin  >> x4;
				cout << "Mate Apr: ";
				cin >> x5;
				cout << "Espa Apr: ";
				cin >> x6;

				y = -12.027 + 2.174*x1 - 0.001*x2 + 0.005*x3 - 0.001*x4 +
0.002*x5 + 0.001*x6;


				if (y >= 0.5 ) // Si r es mayor o igual que 0.5 el
estudiante tendra exito.
				   {
```

```
                                    cout << "\nSaludos estudiante! Por ser un
estudiante tan comprometido y sobresaliente\n"
                                         <<   "en las áreas de las matemáticas eres
un candidato a tener éxito en tu \n"
                                         << "primer curso de matemáticas. Solo te
recomendamos que repases  los conceptos \n"
                                         << "y destrezas básicas para el curso de
matemáticas que estas próximo a comenzar.\n";
                                    }
                               else // Si r es menor que 0.5 el estudiante no tendra
exito.
                                   {

                                     cout << "\nSaludos estudiante! Debes repasar los
conceptos y destrezas \n"
                                         << "básicas para los cursos de matemáticas
que estas próximo a comenzar.\n"
                                         << "De lo contrario puede verse afectado
tu desempeño en el curso. \n";

                                   }
                               break;
                        }

                        case 2: // Calcula la funcion de regresion logistica
                        {
                               double temp1,
                               temp2;

                               cout << "Entre los coeficientes para la funcion de
regresion logistica:"<<endl;
                               cout << "Promedio de Escuela: ";
                               cin  >> x1;
                               cout << "Verbal Apt: ";
                               cin  >> x2;
                               cout << "Mate Apt: ";
                               cin  >> x3;
                               cout << "Ingles Apr: ";
                               cin  >> x4;
                               cout << "Mate Apr: ";
                               cin >> x5;
                               cout << "Espa Apr: ";
                               cin >> x6;

                               temp1 = exp(-6.741 + 1.291*x1 - 0.001*x2 +  0.003*x3 +
0.001*x5 + 0.001*x6);
                               temp2 = 1 + temp1;

                               y = temp1/temp2;

                               cout << exito_fracaso(y);
                               break;
                        }
```

```
                    case 3: //Calcula la funcion logistica
                    {
                           double temp3,
                           temp4;

                           cout << "Entre los coeficientes para la funcion
logistica:"<<endl;
                           cout << "Promedio de Escuela: ";
                           cin  >> x1;
                           cout << "Tipo de escuela(1): ";
                           cin  >> x2;
                           cout << "Verbal Apt: ";
                           cin  >> x3;
                           cout << "Mate Apt: ";
                           cin  >> x4;
                           cout << "Ingles Apr: ";
                           cin >> x5;
                           cout << "Mate Apr: ";
                           cin >> x6;
                           cout << "Espa Apr: ";
                           cin >> x7;



                           temp3 = exp(-7.835 + 1.475*x1 + 0.202*x2 - 0.001*x3 +
0.003*x4 - 0.001*x5 + 0.001*x6 + 0.001*x7) ;
                           temp4 = 1 + temp3;

                           y = temp3/temp4;

                           cout << exito_fracaso(y);
                           break;
                    }
                    case 4:
                    {
                           cout << "Fin del programa" << endl;
                           break;
                    }
                    default:
                    {
                        cout <<" Error en la seleccion del menu."<<endl;
                    }
             } //Fin switch-case
      } while (x != 4 ); //Fin do-while
    return 0;
} //fin del programa
```