

**ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL**



**FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICAS  
DEPARTAMENTO DE MATEMÁTICAS**

**PROYECTO DE TITULACIÓN**

PREVIO A LA OBTENCIÓN DEL TÍTULO DE:

**“MAGÍSTER EN ESTADÍSTICA CON MENCIÓN EN GESTIÓN DE  
LA CALIDAD Y LA PRODUCTIVIDAD”**

**TEMA:**

DESARROLLO DE UN MODELO PARA PREDECIR LA  
PROBABILIDAD DE DESERCIÓN EN LOS ESTUDIANTES DE  
PRIMER AÑO DE UNA INSTITUCIÓN DE EDUCACIÓN SUPERIOR

**AUTOR:**

KAREN STEFANÍE BUSTAMANTE GRACIA

Guayaquil - Ecuador

2018

## RESUMEN

Este trabajo aborda la problemática de la deserción universitaria, el objetivo principal fue predecir la probabilidad de desertar de un estudiante de carreras del área STEM (Ciencia, Tecnología, Ingeniería y Matemática) en una Institución de Educación Superior pública ecuatoriana, a partir de un conjunto de variables personales, académicas y socioeconómicas del estudiante, mediante la aplicación de técnicas de análisis bivariado, y de la implementación de un modelo de regresión logística. La población de estudio estuvo conformada por 371 estudiantes de carreras del área STEM que iniciaron sus estudios en el segundo término académico del año 2011. Se obtuvo que el comenzar la carrera a una edad mayor, el bajo nivel de escolaridad de los padres, un mayor número de miembros en el hogar y reprobar materias núcleo del ciclo básico común de ingenierías tales como Cálculo Diferencial, Física A y Química General, se asocian con una mayor propensión a desertar de la universidad. La evaluación y aplicación de los resultados obtenidos en este estudio permitirá soportar la toma de decisiones, así como orientar la formulación de políticas, lineamientos y estrategias enfocados en reducir la deserción estudiantil universitaria.

Palabras claves: regresión logística, deserción universitaria

## **ABSTRACT**

This project shows the problem of university dropouts. The main objective was to predict the student dropout probability in STEM (Science, Technology, Engineering and Mathematics) careers, of an Ecuadorian public Higher Education Institution, based on a set of personal, academic and socioeconomic variables of the student, through the application of bivariate analysis techniques, and the implementation of a logistic regression model. The study population consisted of 371 STEM students who began their studies in the second academic term of 2011. The results showed that beginning the career at a higher age, fathers with low schooling levels, greater number of members in the home and failing subjects of the common core cycle of engineering such as Differential Calculus, Physics A and General Chemistry, are associated with a greater propensity to drop out university. The evaluation and application of the results obtained in this study will make it possible to support decision making, as well as to guide the formulation of policies, guidelines and strategies focused on reducing university student dropout.

Keywords: Logistic regression, university dropout

## **DEDICATORIA**

*A mi madre y hermanos.*

## **AGRADECIMIENTO**

*A Dios por estar siempre conmigo, a mi familia por su constante apoyo, a todas las personas que de alguna forma colaboraron en la realización de este trabajo, y a ESPOL, mi alma máter.*

## DECLARACIÓN EXPRESA

La responsabilidad por los hechos y doctrinas expuestas en este Proyecto de Graduación, me corresponde exclusivamente; el patrimonio intelectual del mismo, corresponde exclusivamente a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL.



Ing. Karen Stefaníe Bustamante Gracia

## TRIBUNAL DE GRADUACIÓN



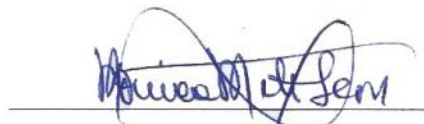
Mgtr. Wendy Plata Alarcón  
Presidente



Mgtr. Mario Solórzano Carvajal  
Director



Ph.D. Sandra García Bustos  
Vocal 1



Mgtr. Mónica Mite León  
Vocal 2

## **ABREVIATURAS O SIGLAS**

IES: Instituciones de Educación Superior

STEM: Ciencia, Tecnología, Ingeniería y Matemática

SNIESE: Sistema Nacional de Información de Educación Superior del Ecuador

SENESCYT: Secretaría de Educación Superior, Ciencia, Tecnología e Innovación

OCDE: Organización para la Cooperación y el Desarrollo Económico

PIB: Producto Interno Bruto

CACES: Consejo de Aseguramiento de la Calidad de la Educación Superior

INSOEC: Encuesta socioeconómica aplicada a los estudiantes en la Institución de Educación Superior de estudio

LOES: Ley Orgánica de Educación Superior

GMS: Gasto Mensual respecto al Salario Mínimo Vital

ROC: Receiver Operating Characteristic, o Característica Operativa del Receptor

AUC: Área bajo la curva



# TABLA DE CONTENIDO

	Página
CAPÍTULO 1 .....	1
1. INTRODUCCIÓN .....	1
1.1. Antecedentes .....	1
1.2. Descripción del problema.....	3
1.3. Objetivos .....	5
1.3.1 Objetivo General.....	5
1.3.2 Objetivos Específicos .....	5
1.4. Alcance .....	6
CAPÍTULO 2 .....	7
2. MARCO TEÓRICO, ESTADO DEL ARTE .....	7
2.1. Marco Teórico .....	7
2.1.1 Tablas de Contingencia y prueba Ji - cuadrado .....	7
2.1.2 Prueba de Levene .....	7
2.1.3 Prueba t para muestras independientes.....	8
2.1.4 Regresión Logística.....	9
2.2. Estado del Arte .....	12
CAPÍTULO 3 .....	14
3. METODOLOGÍA .....	14
3.1. Datos.....	14
3.2. Principales estadísticas descriptivas.....	17
3.2.1 Variable: Edad.....	17
3.2.2 Variable: Sexo .....	18
3.2.3 Variable: Ingresos.....	18
3.2.4 Variable: Porcentaje a pagar por crédito .....	19

3.2.5	Variable: Formación del padre.....	19
3.2.6	Variable: Formación de la madre.....	20
3.2.7	Variable: Miembros Hogar .....	20
3.2.8	Variable: Tipo Colegio .....	21
3.2.9	Variable: Reprobó materias .....	21
3.2.10	Variable: Promedio General .....	22
3.2.11	Variable: Aprobó Cálculo.....	22
3.2.12	Variable: Aprobó Física .....	22
3.2.13	Variable: Aprobó Química General.....	23
3.2.14	Variable Desertó.....	23
CAPÍTULO 4 .....		25
4.	RESULTADOS.....	25
4.1.	Introducción .....	25
4.2.	Asociaciones Bivariantes .....	25
4.2.1	Desertor vs. Edad .....	25
4.2.2	Desertor vs. Sexo.....	26
4.2.3	Desertor vs. Ingresos .....	27
4.2.4	Desertor vs. Porcentaje a Pagar por Crédito .....	27
4.2.5	Desertor vs. Formación del Padre.....	28
4.2.6	Desertor vs. Formación de la Madre .....	28
4.2.7	Desertor vs. Número de Miembros del Hogar.....	29
4.2.8	Desertor vs. Tipo de Colegio.....	29
4.2.9	Desertor vs. Reprobó Materias .....	30
4.2.10	Desertor vs Promedio General.....	30
4.2.11	Desertor vs Aprobó Cálculo .....	31
4.2.12	Desertor vs. Aprobó Física A .....	31
4.2.13	Desertor vs Aprobó Química General .....	32

4.3. Regresión Logística .....	32
4.3.1 Resultados del modelo.....	33
4.3.2 Ajuste del modelo .....	39
CAPÍTULO 5 .....	42
5. CONCLUSIONES Y RECOMENDACIONES .....	42
5.1 CONCLUSIONES.....	42
5.2 RECOMENDACIONES.....	43
6. REFERENCIAS .....	45
7. ANEXOS.....	47
Anexo A. Código de R .....	47

## LISTADO DE FIGURAS

	Página
Figura 4.1 Coeficientes estimados del modelo.....	35
Figura 4.2 Efectos de los factores del modelo sobre la probabilidad de desertar .	38
Figura 4.3 Curva ROC .....	41

## LISTADO DE TABLAS

	Página
Tabla 1.1 Distribución por deciles del acceso a la universidad .....	4
Tabla 3.1 Codificación de la variable dependiente.....	16
Tabla 3.2 Descripción de las variables explicativas .....	16
Tabla 3.3 Estadísticos descriptivos. Variable Edad.....	17
Tabla 3.4 Estadísticos descriptivos. Variable Miembros Hogar.....	21
Tabla 3.5 Estadísticos descriptivos. Variable Promedio General .....	22
Tabla 4.1 Prueba de Levene. Variable Edad .....	26
Tabla 4.2 Prueba t de student. Variable Edad .....	26
Tabla 4.3 Tabla de contingencia. Variables Desertor vs. Sexo .....	26
Tabla 4.4 Tabla de Contingencia. Variables Desertor vs. Ingresos.....	27
Tabla 4.5 Tabla de contingencia. Variables Desertor vs. Porcentaje a Pagar por Crédito .....	27
Tabla 4.6 Tabla de contingencia. Variables Desertor vs. Formación del Padre ....	28
Tabla 4.7 Tabla de contingencia. Variables Desertor vs. Formación de la Madre	28
Tabla 4.8 Prueba de Levene. Variable Miembros Hogar .....	29
Tabla 4.9 Prueba t de student. Variable Miembros Hogar .....	29
Tabla 4.10 Tabla de contingencia. Variables Desertor vs. Tipo de Colegio .....	29
Tabla 4.11 Tabla de contingencia. Variables Desertor vs. Reprobó Materias.....	30
Tabla 4.12 Prueba de Levene. Variable Promedio General .....	30
Tabla 4.13 Prueba t de student. Variable Promedio General .....	30
Tabla 4.14 Tabla de contingencia. Variables Desertor vs. Aprobó cálculo .....	31
Tabla 4.15 Tabla de contingencia. Variables Desertor vs. Aprobó física .....	31
Tabla 4.16 Tabla de contingencia. Variables Desertor vs. Aprobó química .....	32
Tabla 4.17 Codificación de variables del modelo de regresión logística .....	33
Tabla 4.18 Coeficientes $B$ estimados del modelo Logit.....	34
Tabla 4.19 Oportunidades relativas .....	36
Tabla 4.20 Medidas de bondad de ajuste del modelo .....	40
Tabla 4.21 Tabla de clasificación .....	40

## LISTADO DE CUADROS

	Página
Cuadro 3.1 Estadísticas descriptivas. Variable Sexo .....	18
Cuadro 3.2 Estadísticas Descriptivas. Variable Ingresos.....	18
Cuadro 3.3 Estadísticas descriptivas. Variable Porcentaje a Pagar por Crédito ...	19
Cuadro 3.4 Estadísticas descriptivas. Variable Formación del Padre .....	20
Cuadro 3.5 Estadísticas descriptivas. Variable Formación de la Madre .....	20
Cuadro 3.6 Estadísticos descriptivos. Variable Tipo de Colegio .....	21
Cuadro 3.7 Estadísticos descriptivos. Variable Reprobó Materias.....	21
Cuadro 3.8 Estadísticos descriptivos. Variable Aprobó Cálculo.....	22
Cuadro 3.9 Estadísticas descriptivas. Variable Aprobó Física .....	23
Cuadro 3.10 Estadísticas descriptivas. Variable Aprobó Química .....	23
Cuadro 3.11 Estadísticas descriptivas. Variable Desertó.....	24

# CAPÍTULO 1

## 1. INTRODUCCIÓN

### 1.1. Antecedentes

En las últimas décadas América Latina ha dado grandes saltos en el ámbito de la educación superior. Gracias a la preocupación de los Gobiernos por expandir el sistema educativo, ha sido posible aumentar el acceso para los estudiantes a través de la apertura de nuevas instituciones de educación superior (IES), de la diversificación de la oferta académica, del fomento de préstamos y programas de becas, y de políticas sociales cuyo objetivo es proteger a los individuos en los quintiles de distribución de ingresos más bajos. Todo esto permitió incrementar la matrícula en la educación superior, al tiempo que hizo posible que los grupos sociales con bajos ingresos económicos se incorporen al sistema universitario (Rama, La tendencia a la masificación de la cobertura de la educación superior en América Latina, 2009); lo que a su vez ha ocasionado un panorama complejo caracterizado por bajos niveles de aprendizaje y reducida eficiencia terminal (Rama, La educación superior en América Latina ante los nuevos desafíos del ciclo económico, 2016) (CINDA, 2006).

La tasa bruta de matrícula de educación superior en América Latina, indicador que mide el número de estudiantes matriculados en un determinado nivel de enseñanza, independientemente de la edad, expresado en porcentaje de la población en edad escolar oficial correspondiente al mismo nivel de enseñanza, pasó del 29% en el 2004 al 44% en el 2014 (NU. CEPAL, CAF, & OCDE, 2016). Por otro lado, según informe presentado por el Banco Mundial (Marta, Avitabile, Botero, Haimovich, & Urzúa, 2017), la tasa de acceso, indicador que mide la fracción de individuos con edades entre 18 y 24 años que alguna vez han tenido acceso a la educación superior pasó del 18% al 28% del 2000 al 2013; esto en razón del incremento del número de graduados en la secundaria y de las tasas de ingreso a la universidad.

A pesar de las alentadoras cifras de acceso, existen temas que generan preocupación en el sistema de educación superior, tal es el caso de la deserción estudiantil, puesto que de acuerdo al Grupo Banco Mundial, en promedio, la mitad de la población con edades comprendidas entre los 25 y 29 años no finalizó sus estudios universitarios (Marta, Avitabile, Botero, Haimovich, & Urzúa, 2017).

Según la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT), al 2017 el sistema educativo de Ecuador cuenta con 60 universidades y escuelas politécnicas y 278 institutos superiores, técnicos y tecnológicos (SENESCYT, 2017).

De acuerdo a la rendición de cuentas 2016 realizada por este mismo organismo, al 2016 la inversión realizada en educación superior por el Ecuador (2% del PIB) es superior a la realizada por países tales como Bélgica, Francia, Chile, Brasil, Finlandia, E.E.U.U., México, España, entre otros países, esto y las políticas públicas implementadas en los últimos años han permitido incrementar el acceso al sistema universitario. Según cifras de este mismo organismo, el número de matriculados en la educación superior en Ecuador pasó de 443 509 en el 2006 a 759 050 en el 2016, mientras que la tasa bruta de matrícula en estos mismos periodos pasó del 28% al 41%, lo cual representa un incremento de 13 puntos porcentuales (SENESCYT, 2016).

Con respecto al nivel socioeconómico de los matriculados en las IES públicas ecuatorianas, existe una reducción del 65% en la brecha entre el decil más rico y el decil más pobre de matriculados entre los años 2006 y 2014, por otro lado, al 2016 uno de cada dos ecuatorianos que asiste a la universidad proviene de familias en las que los padres no accedieron a la educación superior (Ramírez, 2016).

En lo que refiere a titulación, de acuerdo a la SENESCYT en el 2012 la tasa de titulación fue del 14%, valor que representa el porcentaje de estudiantes que iniciaron estudios universitarios en el 2005 y se titularon hasta el 2012. De acuerdo a cifras de este mismo organismo se espera que este indicador alcance el 58% en el 2018 (SENESCYT, 2016).



A pesar de la importancia de este tema, en el país es muy limitada y escasa la información sobre deserción universitaria; sin embargo, se conoce por estudios realizados por la SENESCYT que el 26% de los estudiantes abandona sus carreras, especialmente en los primeros semestres; cifra que según los representantes gubernamentales se ha logrado disminuir con respecto al año 2010, año en que el porcentaje de desertores alcanzaba el 50% ( Calderón Defaz , Espinel Armas, Garzón Carrera, & Pástor Guevara, 2017).

## **1.2. Descripción del problema**

Se entiende como deserción estudiantil al abandono temporal o definitivo de los estudios (Carvajal, Trejos, & Caro, 2006); la deserción viene dada por diversas causas, académicas y no académicas, y ocurre con mayor frecuencia en los primeros niveles de estudio.

La deserción produce varios problemas en los países, tanto sociales como económicos, uno de los principales es el costo de la educación superior por estudiante (CINDA, 2006), financiado por el Estado en el caso de las instituciones públicas, condicionando de esta manera el desarrollo de los países, por el desaprovechamiento de los recursos fiscales.

Los países de la Organización para la Cooperación y el Desarrollo Económico (OCDE) gastan en promedio USD 15 028 al año por estudiante de tercer nivel (NU. CEPAL, CAF, & OCDE, 2016). El promedio de gasto por estudiante en América Latina es comparable al de Europa central y oriental (NU. CEPAL, CAF, & OCDE, 2016). Si medimos el gasto por estudiante como porcentaje del producto interior bruto (PIB), el gasto por estudiante es bastante similar al de los países desarrollados, lo cual es prueba del esfuerzo realizado por la región pese a las diferencias económicas de las regiones (NU. CEPAL, CAF, & OCDE, 2016).

La deserción repercute también de forma negativa en la economía de las familias de escasos recursos, debido a que los padres realizan grandes esfuerzos para lograr que sus hijos ingresen a la universidad, en especial aquellos que envían a sus hijos a estudiar fuera de su ciudad de residencia. Esta situación es muy común

en nuestro país, en especial a partir de la implementación del actual sistema de nivelación y admisión unificado.

Existen además un conjunto de repercusiones psicológicas y afectivas en el estudiante al no lograr la consecución de sus metas, lo cual a futuro también disminuye sus oportunidades de conseguir un empleo favorable.

Por otro lado, en la actualidad las universidades se enfrentan a bajos índices de titulación producto de la masificación del ingreso y del acceso de los sectores históricamente excluidos al sistema de educación superior, esto se puede evidenciar al analizar la distribución por deciles del acceso a la universidad (tabla 1.1) donde se puede notar que los cinco deciles más pobres de la distribución, representaban apenas el 12% de la matrícula universitaria en el 2006, cifra que alcanzó el 18% en el 2014; mientras que el decil con mayores ingresos pasó del 32% de participación en la distribución total en el 2006, al 25% en el 2014.

**Tabla 1.1 Distribución por deciles del acceso a la universidad**

Decil	Distribución total		Distribución acumulada	
	2006	2014	2006	2014
10% más pobre	0,2%	1,1%	0,2%	1,1%
2	1,1%	2,1%	1,3%	3,1%
3	1,7%	2,8%	3,0%	5,9%
4	2,8%	5,5%	5,8%	11,5%
5	6,0%	6,3%	11,8%	17,8%
6	6,5%	7,8%	18,3%	25,6%
7	10,8%	12,2%	29,1%	37,8%
8	16,2%	15,6%	45,4%	53,3%
9	22,7%	21,4%	68,1%	74,8%
10% más rico	31,9%	25,2%	100,0%	100,0%

*Nota.* Recuperado de “Acceso y equidad a la educación superior y posgrado en el Ecuador, un enfoque descriptivo”, de Ponce, Juan.; Carrasco, Fernando, 2016, *Mundos Plurales*, Vol. 3, p. 9 - 22. Copyright 2016 de FLACSO.

El número de desertores afecta directamente la tasa de titulación, indicador que mide el porcentaje de estudiantes que se titulan dentro del tiempo estimado de duración de su programa académico; este es considerado como uno de los principales indicadores de eficiencia académica, y sus resultados forman parte del modelo de evaluación para las instituciones de educación superior (IES) ecuatorianas, propuesto por el Consejo de Aseguramiento de la Calidad de la Educación Superior (CACES), por lo que las universidades se encuentran en la

búsqueda permanente de estrategias y programas que garanticen la permanencia de los estudiantes hasta la finalización de sus estudios.

Otra implicancia directa es la asignación de recursos que reciben las IES públicas por parte del Estado, ya que la tasa de titulación es una componente de la fórmula utilizada por el Consejo de Educación Superior (CES) para la distribución de recursos, pues se considera un parámetro de calidad.

Por lo antes mencionado, es necesario que las IES cuenten con una metodología que permita identificar tempranamente a los estudiantes vulnerables, así como los factores de riesgo y de protección asociados a este fenómeno, con lo cual será posible diseñar y mejorar las estrategias de seguimiento y monitoreo, y por consiguiente disminuir la ocurrencia de la deserción.

### **1.3. Objetivos**

#### **1.3.1 Objetivo General**

Predecir la probabilidad de deserción estudiantil como mecanismo de detección temprana para proporcionar apoyo tutorial integral al estudiante y fortalecer las estrategias de prevención.

#### **1.3.2 Objetivos Específicos**

- Determinar si existe dependencia entre la deserción estudiantil y un conjunto de variables inherentes del estudiante.
- Detectar factores determinantes (de riesgo y de protección) relacionados con la deserción estudiantil.
- Predecir la probabilidad de deserción de un estudiante.

#### **1.4. Alcance**

Este proyecto busca determinar un modelo para predecir la probabilidad de desertar de un estudiante de primer año de una institución de educación superior pública ubicada en la ciudad de Guayaquil, provincia del Guayas, perteneciente a carreras del área STEM (Ciencia, Tecnología, Ingeniería y Matemáticas).

## CAPÍTULO 2

### 2. MARCO TEÓRICO, ESTADO DEL ARTE

#### 2.1. Marco Teórico

Para la realización de estudio se emplearon varias técnicas estadísticas, a continuación se presentan de forma breve los principales conceptos teóricos necesarios para abordar los resultados obtenidos en esta investigación.

##### 2.1.1 Tablas de Contingencia y prueba Ji - cuadrado

Las tablas de contingencia son arreglos rectangulares que permiten determinar si existe alguna relación entre dos variables diferentes de una población, estas variables o características a su vez se encuentran divididas en categorías o criterios.

La tabla se compone de las frecuencias relativas observadas para cada variable con sus correspondientes categorías. La hipótesis nula ( $H_0$ ) a probar es que las dos características son independientes, si  $H_0$  es rechazada indicaría que existe relación entre las variables analizadas (Canavos, 1988).

El estadístico de prueba Ji-cuadrado, compara las frecuencias observadas con las esperadas, para un nivel de significación  $\alpha$  la hipótesis  $H_0$  se acepta si:

$$\sum_{i=1}^k \sum_{j=1}^m \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \leq \chi_{\alpha, (k-1)(m-1)}^2$$

y se rechaza en caso contrario (Canavos, 1988).

##### 2.1.2 Prueba de Levene

La prueba de Levene es un contraste que nos permite analizar la homogeneidad de varianzas. La hipótesis nula  $H_0$  es que existe igualdad de varianza en las

poblaciones, mientras que la hipótesis alterna  $H_1$  es que no existe igualdad de varianzas (Rubio Hurtado & Berlanga Silvente, 2012).

El estadístico que utiliza esta prueba para contrastar las hipótesis es el estadístico F de Snedecor.

$$F = \frac{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}{\frac{1}{m-1} \sum_{i=1}^m (Y_i - \bar{Y})^2} = \frac{\hat{S}_1^2}{\hat{S}_2^2}$$

### 2.1.3 Prueba t para muestras independientes

Esta prueba se aplica cuando los individuos de las poblaciones a comparar son distintos y contrasta si la media de una población difiere significativamente de la media de la otra población. Existen dos formas para calcular el estadístico de esta prueba, que dependen de si se asumen o no varianzas iguales. (Zurita Herrera, 2008).

Cuando dos muestras independientes son extraídas de poblaciones que se asumen tienen varianzas idénticas, el estadístico t se calcula de la siguiente manera:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Con

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Donde:

$\bar{x}_1$  = Media de la primera muestra

$\bar{x}_2$  = Media de la segunda muestra

$n_1$  = Tamaño de la primera muestra

$n_2$  = Tamaño de la segunda muestra

$s_1$  = Desviación estándar de la primera muestra  
 $s_2$  = Desviación estándar de la segunda muestra  
 $s_p$  = Desviación estándar agrupada

Cuando dos muestras independientes son extraídas de poblaciones que se asumen tienen varianzas desiguales, el estadístico t se calcula de la siguiente manera:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Donde:

$\bar{x}_1$  = Media de la primera muestra

$\bar{x}_2$  = Media de la segunda muestra

$n_1$  = Tamaño de la primera muestra

$n_2$  = Tamaño de la segunda muestra

$s_1$  = Desviación estándar de la primera muestra

$s_2$  = Desviación estándar de la segunda muestra

#### 2.1.4 Regresión Logística

La regresión logística es una técnica multivariada que tiene como objetivo explicar una determinada variable dicotómica, que toma el valor de “0” si el evento no ocurre y “1” en caso de que sí, a través de un conjunto de variables que se presumen tienen un efecto sobre la variable dependiente.

Esta técnica también es utilizada para predecir la ocurrencia de un determinado evento en función de un conjunto de factores que se presumen relevantes sobre la variable dependiente, y determinar cómo influyen cada una de las variables explicativas en el aumento o disminución de la probabilidad.

Las variables explicativas pueden ser de tipo cuantitativas o cualitativas, para el caso de las covariables categóricas es necesario que sean transformadas en variables indicadoras “dummy”, que son aquellas que toman el valor de 1 o 0 para

indicar la presencia o ausencia de un evento categórico que tal vez pueda provocar un efecto en el resultado (Everitt & Dunn, 2001).

El modelo de regresión logística se puede representar de la siguiente forma:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$$

Donde:

$p_i$  = probabilidad de que ocurra el evento de interés

$1 - p_i$  = probabilidad de que no ocurra el evento de interés

A través de los coeficientes de regresión  $\beta$  del modelo es posible obtener los *odds ratio*, los cuales reflejan el efecto del aumento o disminución de la variable independiente sobre la variable de estudio, es decir permiten cuantificar la magnitud entre la respuesta y el factor de interés.

La estimación de los coeficientes del modelo se realiza a través del cálculo de estimaciones de máxima verosimilitud, ejecutados a través de algoritmos iterativos como el de Newton Raphson, por lo que es necesario la aplicación de paquetes computacionales para su consecución. El contraste de los coeficientes se lo realiza a través de la prueba de Wald, cuya hipótesis a probar es que un coeficiente aislado es distinto de 0, lo cual indicaría que el modelo es adecuado para representar una determinada relación.

La regresión logística al igual que otros modelos de regresión, compara los valores observados de la variable de respuesta con los variables predichos obtenidos del modelo, en dos momentos: el primero en el modelo inicial, es decir considerando solo la constante, y el segundo al incluir una o más variables; esta comparación se realiza a través de la función de máxima verosimilitud, basada en la siguiente expresión (Hosmer & Lemeshow, 2013):



$$D = -2 \ln \left[ \frac{(\text{probabilidad del modelo ajustado})}{(\text{probabilidad del modelo saturado})} \right]$$

El estadístico  $D$  es llamado desviación, y en el caso de la regresión logística realiza la función de la suma de cuadrados residual en la regresión lineal (Hosmer & Lemeshow, 2013).

Los modelos de regresión logística se ajustan utilizando el método de máxima verosimilitud, es decir, los parámetros estimados son aquellos valores que maximizan la probabilidad de que los resultados se hayan observado (Smith & McKenna, 2013). El  $R^2$  de McFadden, es análogo al  $R^2$  de la regresión lineal, y se define la siguiente manera:

$$R_{McFadden}^2 = 1 - \frac{\log(L_c)}{\log(L_n)}$$

Donde  $L_c$  es la función de verosimilitud del modelo completo, es decir con todas las variables explicativas, y  $L_n$  es la función de verosimilitud del modelo nulo (el modelo solo con el intercepto, sin covariables).

Otra forma de evaluar el modelo es a través de la prueba de Hosmer y Lemeshow, la cual evalúa la predicción real de la variable dependiente, dividiendo los casos en clases y comparando las frecuencias esperadas con las observadas en cada una de ellas (Hosmer & Lemeshow, 2013). El estadístico se define como:

$$c = \sum_{i=1}^g \frac{(O_i - n_i \bar{p}_i)^2}{n_i \bar{p}_i (1 - \bar{p}_i)}$$

Donde:

$g$  es el número de grupos

$n'_i$  es el número de observaciones en el  $i$ -ésimo grupo

$O_i$  es la suma de las  $Y$  en el  $i$ -ésimo grupo;

$\bar{p}_i$  es el promedio de las  $p_i$  en el  $i$ -ésimo grupo

## 2.2. Estado del Arte

A nivel internacional existen investigaciones similares relacionadas con la deserción estudiantil que demuestran que es posible aplicar técnicas de minería de datos tales como la regresión logística para detectar a los estudiantes más vulnerables, con mayor riesgo de desertar.

(Trevizán, Beltrán, & Cosolito, 2009) realizaron un estudio en Argentina para determinar las variables de los estudiantes de nuevo ingreso de la carrera Ingeniería Agronómica que condicionan la deserción durante su trayecto universitario; para ello aplicaron la técnica de regresión logística e identificaron que las características de los alumnos que están asociadas a la deserción universitaria son: el comienzo de la carrera a una edad mayor, padres con bajo nivel educativo y tipo de residencia con padres y hermanos.

Una investigación realizada en estudiantes de la carrera de ingeniería civil de una universidad chilena determinó, a través de la utilización de la regresión logística binaria, que las variables que mejor explican la deserción estudiantil son el rendimiento académico y los ingresos familiares, investigación basada en el modelo de deserción de Vincent Tinto (Seldaña Villa & Barriga, 2010).

Por otro lado, (Roth, 2008) desarrolló varios modelos de regresión lineal, regresión logística y alternancia de decisiones para predecir la deserción de los estudiantes en la Universidad de Ohio. En este trabajo fue posible predecir a un alto número de estudiantes desertores.

A nivel nacional, es muy limitada la bibliografía que documente resultados significativos de la aplicación de acciones al respecto, pero pueden citarse los siguientes trabajos:

(Fernandez & Silva, 2014) a través del uso de la regresión logística presentan las relaciones de un conjunto de variables universitarias con la deserción estudiantil, entre los principales hallazgos se destacan que las notas del colegio y las notas de

ingreso a la universidad no influyen en la deserción, a diferencia de la edad y el número de materias tomadas que resultaron significativas.

Por su parte, (Del Alcázar León, Silva, & Fernández Orrantia, 2014) en su estudio cuantificaron y cualificaron la asociación entre las posibles causas de la deserción mediante red neuronal, para generar un perfil de riesgo enfocado en estudiantes de carreras técnicas. Las variables que permitieron explicar la deserción estudiantil en este estudio fueron: las materias reprobadas en el primer semestre, las materias aprobadas en el primer semestre, la nota del colegio, la nota de admisión, la edad, el número de materias tomadas en el primer semestre, el tipo de colegio y el sexo.

Es común encontrar que en estos trabajos los esfuerzos se centran en los primeros años de vida estudiantil, en el caso de Ecuador por la concurrencia de varios factores interrelacionados tales como el bajo nivel de conocimientos en ciencias básicas adquiridos en la educación secundaria; el desencuentro entre la carrera que anheló y la que se registró el estudiante; y, la transición a la que se enfrentan los estudiantes al tener que adaptarse a diferentes metodologías de enseñanza-aprendizaje, entre otros factores.

Los trabajos citados previamente, demuestran que mediante la aplicación de técnicas de minería de datos es posible pasar de datos a información útil para soportar la toma de decisiones en la gestión institucional de las IES.

## **CAPÍTULO 3**

### **3. METODOLOGÍA**

La metodología empleada para el desarrollo de esta investigación comprende inicialmente la recopilación y depuración de los datos disponibles de los estudiantes, así como el análisis descriptivo de los resultados más relevantes.

La segunda etapa consiste en la elaboración de tablas de contingencia para determinar si existe dependencia entre la deserción estudiantil y las variables relacionadas con el estudiante, esto permitirá seleccionar del conjunto inicial de variables, aquellas que estén asociadas con la deserción.

Posteriormente, a través de un modelo lineal generalizado de regresión logística se explica la variable dependiente binaria que indica si el estudiante desertó o no al menos una vez durante sus primeros cinco años de estudio, a través de variables inherentes al estudiante; y se determinan factores que reducen (de protección) y aumentan (de riesgo) la probabilidad de desertar.

La deserción estudiantil, se define según resolución interna en esta IES como el abandono de la institución por al menos dos términos académicos consecutivos, por cualquier causa, sea esta de índole académica (pérdida de carrera por reprobado una materia por tercera vez) o personal (condiciones socioeconómicas, familiares, entre otras). La variable de respuesta a lo largo de este estudio, indica si el estudiante desertó o no al menos una vez durante los primeros cinco años de su carrera (duración teórica de la carrera).

#### **3.1. Datos**

Para la realización de este estudio se consideraron datos históricos de los 371 estudiantes novatos del segundo término académico del año 2011, pertenecientes a carreras del área STEM, en una institución de educación superior pública ecuatoriana, en el caso de esta institución las carreras del área STEM son las siguientes:

- Biología Marina
- Ingeniería Agrícola y Biológica
- Ingeniería Civil
- Ingeniería de Minas
- Ingeniería de Petróleo
- Ingeniería Electrónica y Telecomunicaciones
- Ingeniería en Acuicultura
- Ingeniería en Alimentos
- Ingeniería en Ciencias Computacionales
- Ingeniería en Electricidad especialización Electrónica y Automatización Industrial
- Ingeniería en Electricidad especialización Potencia
- Ingeniería en Estadística Informática
- Ingeniería en Geología
- Ingeniería en Logística y Transporte
- Ingeniería en Telemática
- Ingeniería Mecánica
- Ingeniería Naval
- Ingeniería Oceánica y Ciencias Ambientales
- Ingeniería Química
- Ingeniería y Administración de la Producción Industrial

La variable de respuesta fue medida durante un periodo de cinco años, esto es desde el segundo término académico del año 2011, hasta el segundo término del año 2016; término en el que, según la duración teórica de estas carreras, dichos estudiantes debieron graduarse.

Los datos fueron obtenidos a través de la Secretaría Técnica Académica de la institución, y para la selección de las variables a utilizar se tomó como referencia estudios previos realizados en otras universidades, y la disponibilidad de estas variables en las bases de datos de la institución.

Las variables socioeconómicas provienen de la encuesta INSOEC realizada por la institución, que tiene como objetivo determinar el porcentaje a pagar por crédito, en cada materia en la que se matricule el estudiante, en caso de que pierda la gratuidad parcialmente.

La variable dependiente en este estudio es la deserción, cuya descripción se muestra en la tabla 3.1.

**Tabla 3.1 Codificación de la variable dependiente**

Variable	Descripción	Codificación original	Recodificación
<i>Desertor</i>	Indica si el estudiante desertó al menos una vez durante los 5 años que duró su carrera	0: Sí 1: No	-

*Fuente: Secretaría Técnica Académica  
Elaboración: Autor*

Las covariables a utilizar se describen en la tabla 3.2.

**Tabla 3.2 Descripción de las variables explicativas**

	Covariables	Descripción	Codificación original	Recodificación
Características del estudiante	<i>Edad_Termino</i>	Años cumplidos por el estudiante en el año-término de estudio	Variable cuantitativa	-
	<i>Sexo</i>	Sexo del estudiante	0: Hombre 1: Mujer	-
Información socioeconómica	<i>Ingresos</i>	Ingreso familiar mensual	Variable cuantitativa	0: Más de \$400 1: Menor o igual a \$400
	<i>Porcentaje_crédito</i>	Porcentaje a pagar por crédito en caso de pérdida de gratuidad parcial	Variable cuantitativa	0: >=75 1: 50.01-75 2: 25.01-50 3: <=25
	<i>Formación_padre</i>	El grado académico más alto al que asiste o asistió el padre del estudiante	0: Postgrado 1: Superior 2: Secundario 3: Primario 4: Ninguno 5: No sabe	0: Mas de secundaria 1: Secundaria o menos 2: No sabe

	<b>Covariables</b>	<b>Descripción</b>	<b>Codificación original</b>	<b>Recodificación</b>
Información académica	<i>Formación_madre</i>	El grado académico más alto al que asiste o asistió la madre del estudiante	0: Postgrado 1: Superior 2: Secundario 3: Primario 4: Ninguno 5: No sabe	0: Mas de secundaria 1: Secundaria o menos 2: No sabe
	<i>Miembros_Hogar</i>	Número de miembros en el hogar	Variable cuantitativa	-
	<i>Tipo_colegio</i>	Tipo de colegio de proveniencia del estudiante	0: Particular 1: Fiscomisional 2: Fiscal	0: Particular 1: Fiscomisional o Fiscal
	<i>Reprobó_materias</i>	Indica si el estudiante reprobó al menos una materia durante su primer término académico	Variable cuantitativa	0: No 1: Sí
	<i>Promedio_General</i>	Promedio general del estudiante finalizado su primer término académico	Variable cuantitativa	-
	<i>Aprobó_Cálculo</i>	Indica si el estudiante aprobó o no la materia de Cálculo Diferencial, al registrarse por primera vez	Variable cuantitativa	0: Sí 1: No
	<i>Aprobó_Física</i>	Indica si el estudiante aprobó o no la materia de Física A, al registrarse por primera vez	Variable cuantitativa	0: Sí 1: No
	<i>Aprobó_Química</i>	Indica si el estudiante aprobó o no la materia de Química General, al registrarse por primera vez	Variable cuantitativa	0: Sí 1: No

*Fuente: Secretaría Técnica Académica  
Elaboración: Autor*

### 3.2. Principales estadísticas descriptivas

#### 3.2.1 Variable: Edad

Los estadísticos descriptivos de la tabla 3.3, muestran que la variable *Edad* toma un valor mínimo de 17, un valor máximo de 24, con una media de 18 y una desviación estándar de 0.974.

**Tabla 3.3 Estadísticos descriptivos. Variable Edad**

	<b>Estadísticos descriptivos</b>				
	N	Mínimo	Máximo	Media	Desv.
<i>Edad_Termino</i>	371	17	24	18	0.974

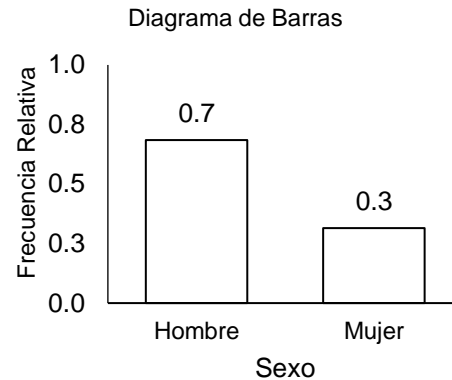
*Fuente: Secretaría Técnica Académica  
Elaboración: Autor*

### 3.2.2 Variable: Sexo

En el cuadro 3.1 se muestra que el 70% de los estudiantes son hombres, mientras que el 30% restante son mujeres.

**Cuadro 3.1 Estadísticas descriptivas. Variable Sexo**

Sexo	Frecuencia Absoluta	Frecuencia Relativa
Mujer	117	0.3
Hombre	254	0.7
Total	371	1.0



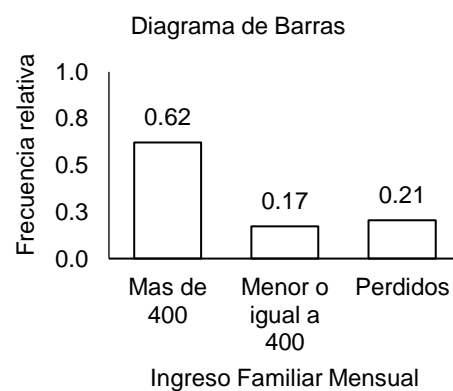
Fuente: Secretaría Técnica Académica  
Elaboración: Autor

### 3.2.3 Variable: Ingresos

En el cuadro 3.2 se muestra que el 62% de los estudiantes indicó que su ingreso familiar mensual es superior a \$400, mientras que el 17% indicó que este valor es menor o igual a los \$400. El 21% restante no respondió esta pregunta.

**Cuadro 3.2 Estadísticas Descriptivas. Variable Ingresos**

Ingresos	Frecuencia Absoluta	Frecuencia Relativa
Más de 400	231	0.62
Menor o igual a 400	64	0.17
Perdidos	76	0.21
Total	371	1.0



Fuente: Secretaría Técnica Académica  
Elaboración: Autor

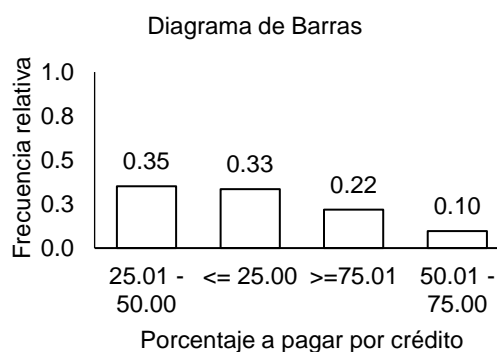


### 3.2.4 Variable: Porcentaje a pagar por crédito

En concordancia con la Ley Orgánica de Educación Superior (LOES), en la IES de estudio existe un valor a pagar por crédito para cada materia en la que se matricule el estudiante, en caso de que pierda la gratuidad parcialmente, la determinación del arancel se compone de dos partes, la primera a partir del quintil  $p$  del Gasto Mensual respecto al Salario Mínimo Vital (GMS)  $Q_p$  calculado a partir de la encuesta socioeconómica, y otra proporcional a partir del historial de reprobación del estudiante. El cuadro 3.3 muestra que el 22% tiene un porcentaje a pagar por crédito superior al 75%, el 10% entre 50.01% y 75%, el 35% entre 25.01% y 50%, y el 33% restante menor igual a 25%.

**Cuadro 3.3 Estadísticas descriptivas. Variable Porcentaje a Pagar por Crédito**

Porcentaje a pagar por crédito	Frecuencia absoluta	Frecuencia Relativa
$\geq 75.01$	81	0.22
50.01 – 75.00	36	0.10
25.01 – 50.00	130	0.35
$\leq 25.00$	124	0.33
Total	371	1.0



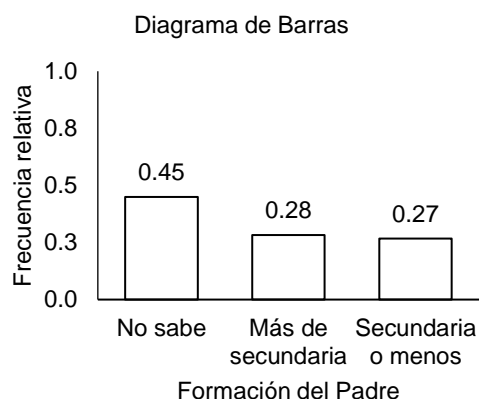
Fuente: Secretaría Técnica Académica  
Elaboración: Autor

### 3.2.5 Variable: Formación del padre

En el cuadro 3.4 se puede observar que el 28% de los padres de los estudiantes cuentan con instrucción de tercer nivel o de postgrado, el 27% secundaria o menos de secundaria, mientras que el 45% restante indicó que no sabe.

**Cuadro 3.4 Estadísticas descriptivas. Variable Formación del Padre**

Nivel de Formación del Padre	Frecuencia	Frecuencia relativa
No sabe	167	0.45
Más de secundaria	105	0.28
Secundaria o menos	99	0.27
Total	371	1.0



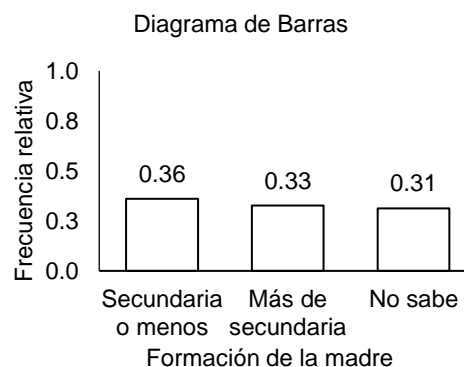
Fuente: Secretaría Técnica Académica  
Elaboración: Autor

### 3.2.6 Variable: Formación de la madre

En el cuadro 3.5 se muestra que el 33% de las madres de los estudiantes cuentan con instrucción de tercer nivel o de postgrado, el 36% secundaria o menos de secundaria, mientras que el 31% restante indicó que no sabe.

**Cuadro 3.5 Estadísticas descriptivas. Variable Formación de la Madre**

Nivel de Formación de la Madre	Frecuencia	Frecuencia relativa
Secundaria o menos	134	0.36
Más de secundaria	121	0.33
No sabe	116	0.31
Total	371	1.0



Fuente: Secretaría Técnica Académica  
Elaboración: Autor

### 3.2.7 Variable: Miembros Hogar

Los estadísticos descriptivos de la tabla 3.4, muestran que la variable número de miembros en el hogar toma un valor mínimo de 1, un valor máximo de 8, con una media aproximada de 3 y una desviación estándar de 1.62.

**Tabla 3.4 Estadísticos descriptivos. Variable Miembros Hogar**

	Estadísticos descriptivos				
	N	Mínimo	Máximo	Media	Desv.
<i>Miembros_Hogar</i>	371	1	8	2.76	1.62

Fuente: Secretaría Técnica Académica

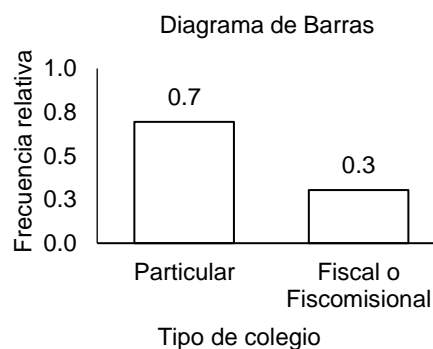
Elaboración: Autor

### 3.2.8 Variable: Tipo Colegio

En el cuadro 3.6 podemos observar que el 70% de los estudiantes proviene de colegios particulares mientras que el 30% restante de colegios fiscales o fiscomisionales.

**Cuadro 3.6 Estadísticos descriptivos. Variable Tipo de Colegio**

Tipo de Colegio	Frecuencia	Frecuencia relativa
Particular	258	0.7
Fiscal o Fiscomisional	113	0.3
Total	371	1.0



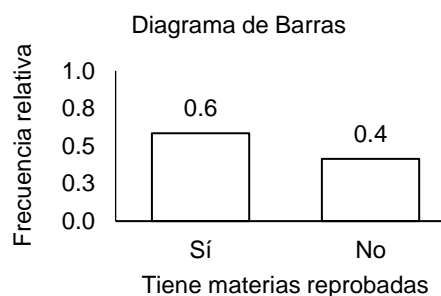
Fuente: Secretaría Técnica Académica

### 3.2.9 Variable: Reprobó materias

El cuadro 3.7 muestra que el 60% de los estudiantes reprobó al menos una materia en su primer término académico, mientras que el 40% restante no reprobó materias.

**Cuadro 3.7 Estadísticos descriptivos. Variable Reprobó Materias**

Reprobó materias	Frecuencia	Frecuencia relativa
Sí	217	0.6
No	154	0.4
Total	371	1.0



Fuente: Secretaría Técnica Académica

Elaboración: Autor

### 3.2.10 Variable: Promedio General

Los estadísticos descriptivos de la tabla 3.5, muestran que la variable promedio general toma un valor mínimo de 0, un valor máximo de 10, con una media de 7.07 y una desviación estándar de 1.15.

**Tabla 3.5 Estadísticos descriptivos. Variable Promedio General**

	Estadísticos descriptivos				
	N	Mínimo	Máximo	Media	Desv.
<i>Promedio General</i>	371	0	10	7.07	1.15

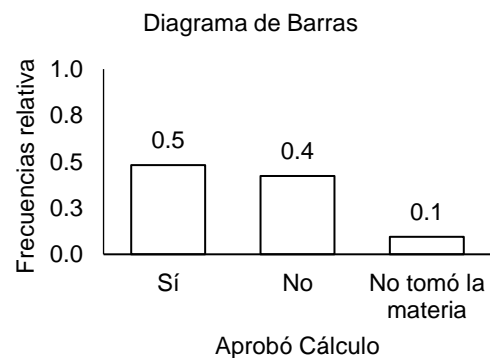
*Fuente: Secretaría Técnica Académica  
Elaboración: Autor*

### 3.2.11 Variable: Aprobó Cálculo

El cuadro 3.8 permite observar que el 50% de los estudiantes aprobó Cálculo Diferencial al tomar la materia por primera vez, mientras que el 40% reprobó; el 10% restante no tomó la materia durante su primer año.

**Cuadro 3.8 Estadísticos descriptivos. Variable Aprobó Cálculo**

Aprobó cálculo	Frecuencia	Frecuencia relativa
Sí	179	0.5
No	157	0.4
No tomó materia	35	0.1
Total	371	1.0



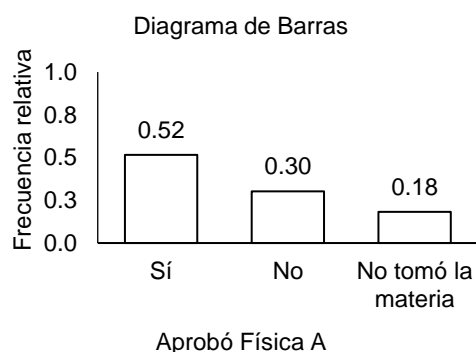
*Fuente: Secretaría Técnica Académica  
Elaboración: Autor*

### 3.2.12 Variable: Aprobó Física

El cuadro 3.9 muestra que el 52% de los estudiantes aprobó Física A, al tomar la materia por primera vez, mientras que el 30% reprobó; el 18% restante no tomó la materia durante su primer año.

**Cuadro 3.9 Estadísticas descriptivas. Variable Aprobó Física**

Aprobó Física A	Frecuencia	Frecuencia relativa
Sí	191	0.52
No	112	0.30
No tomó la materia	68	0.18
Total	371	1.0



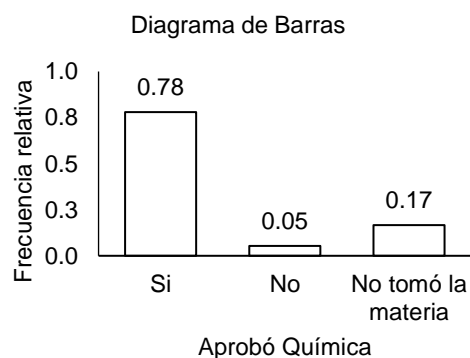
Fuente: Secretaría Técnica Académica  
Elaboración: Autor

### 3.2.13 Variable: Aprobó Química General

El cuadro 3.10 permite observar que el 78% de los estudiantes aprobó Química General al tomar la materia por primera vez, mientras que el 5% reprobó; el 17% restante no tomó la materia durante su primer año.

**Cuadro 3.10 Estadísticas descriptivas. Variable Aprobó Química**

Aprobó Química	Frecuencia	Frecuencia relativa
Sí	289	0.78
No	20	0.05
No tomó la materia	62	0.17
Total	371	1.0



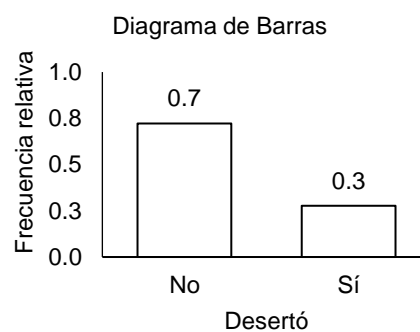
Fuente: Secretaría Técnica Académica  
Elaboración: Autor

### 3.2.14 Variable Desertó

El cuadro 3.11 muestra que el 70% de los estudiantes no desertó, mientras que el 30% restante sí desertó al menos una vez durante sus primeros cinco años de estudio.

**Cuadro 3.11 Estadísticas descriptivas. Variable Desertó**

<b>Desertó</b>	<b>Frecuencia</b>	<b>Frecuencia relativa</b>
No	268	0.7
Sí	103	0.3
Total	371	1.0



*Fuente: Secretaría Técnica Académica  
Elaboración: Autor*

# CAPÍTULO 4

## 4. RESULTADOS

### 4.1. Introducción

En este capítulo se muestran los resultados obtenidos al aplicar las técnicas mencionadas en capítulos previos. La exploración de las asociaciones bivariantes se realizó a través de tablas de contingencia y pruebas ji-cuadrado para el caso de las variables independientes nominales, mientras que para el caso de las variables independientes cuantitativas el proceso empleado fueron las pruebas t de student. Estos análisis nos permiten tener una primera aproximación a la estimación a la medida de asociación entre las variables explicativas y la variable que buscamos predecir.

Posteriormente se muestran los resultados de la aplicación de la regresión logística multivariante, a través de la cual fue posible explicar la deserción en función de un conjunto de variables personales, socioeconómicas y académicas del estudiante, así como también calcular la probabilidad de que un estudiante perteneciente al área STEM deserte durante los primeros cinco años de su carrera universitaria en una institución de educación superior.

### 4.2. Asociaciones Bivariantes

En este apartado se realizan diferentes pruebas estadísticas que permiten contrastar las hipótesis planteadas. Se fijó un nivel de significación del 5%.

#### 4.2.1 Desertor vs. Edad

Al realizar la prueba de Levene para igualdad de varianzas (tabla 4.1) se obtuvo un valor p menor a 0.05 por lo que no es posible asumir que las varianzas son iguales; por esta razón suponemos que las varianzas no son iguales, con lo que al realizar la prueba t de student para la igualdad de medias se obtiene un valor p menor a

0.05 lo que indica que sí existen diferencias significativas en las medias de las edades de los desertores vs. no desertores.

**Tabla 4.1 Prueba de Levene. Variable Edad**

	F	Sig.
<i>Edad_Término</i>	15.236	0.000

*Fuente: Secretaría Técnica Académica  
Elaboración: Autor*

**Tabla 4.2 Prueba t de student. Variable Edad**

	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
						Inferior	Superior
<i>Edad_Término</i>	-2.123	138.246	0.036	-0.283	0.133	-0.547	-0.019

*Fuente: Secretaría Técnica Académica  
Elaboración: Autor*

#### 4.2.2 Desertor vs. Sexo

La tabla 4.3 muestra el valor p de la prueba de independencia entre las variables, el valor p resultante es menor a 0.05, por lo tanto se concluye que estas variables son dependientes entre sí.

**Tabla 4.3 Tabla de contingencia. Variables Desertor vs. Sexo**

Sexo	No Desertor	Desertor	Total
Masculino	172 64.2%	82 79.6%	254 68.5%
Femenino	96 35.8%	21 20.4%	117 31.5%
Total	268 100.0%	103 100.0%	371 100.0%
Ji-cuadrado = 8.27		g.l. = 1	p= 0.004

*Fuente: Secretaría Técnica Académica  
Elaboración: Autor*



### 4.2.3 Desertor vs. Ingresos

La tabla 4.4 muestra el valor p de la prueba de independencia entre las variables, el valor p resultante es mayor a 0.05, por lo tanto se concluye que estas variables son independientes entre sí.

**Tabla 4.4 Tabla de Contingencia. Variables Desertor vs. Ingresos**

Ingreso familiar mensual	No Desertor	Desertor	Total
Más de \$400	199 78.3%	32 78.0%	231 78.3%
Menor igual que \$400	55 21.7%	9 22.0%	64 21.7%
Total	254 100.0%	41 100.0%	295 100.0%
Ji-cuadrado = 0.002		g.l. = 1	p= 0.966

Fuente: Secretaría Técnica Académica  
Elaboración: Autor

### 4.2.4 Desertor vs. Porcentaje a Pagar por Crédito

La tabla 4.5 muestra el valor p de la prueba de independencia entre las variables, el valor p resultante es menor a 0.05, por lo tanto se concluye que estas variables son dependientes entre sí.

**Tabla 4.5 Tabla de contingencia. Variables Desertor vs. Porcentaje a Pagar por Crédito**

Porcentaje a Pagar por Crédito	No Desertor	Desertor	Total
7>=5.01	19 7.1%	62 60.2%	81 21.8%
50.01 – 75.00	32 11.9%	4 3.9%	36 9.7%
25.01 – 50.00	107 39.9%	23 22.3%	130 35.0%
<= 25.00	110 41.0%	14 13.6%	124 33.4%
Total	268 100.0%	103 100.0%	371 100.0%
Ji-cuadrado = 124.434		g.l. = 3	p= 0.000

Fuente: Secretaría Técnica Académica  
Elaboración: Autor

#### 4.2.5 Desertor vs. Formación del Padre

La tabla 4.6 muestra el valor p de la prueba de independencia entre las variables, el valor p resultante es menor que 0.05, por lo tanto se concluye que estas variables son dependientes entre sí.

**Tabla 4.6 Tabla de contingencia. Variables Desertor vs. Formación del Padre**

Nivel de Formación del Padre	No Desertor	Desertor	Total
Más de secundaria	99 36.9%	6 5.8%	105 28.3%
Secundaria o menos	83 31.0%	16 15.5%	99 26.7%
No sabe	86 32.1%	81 78.6%	167 45.0%
Total	268 100.0%	103 100.0%	371 100.0%

Ji-cuadrado = 67.915

g.l. = 2

p= 0.000

Fuente: Secretaría Técnica Académica  
Elaboración: Autor

#### 4.2.6 Desertor vs. Formación de la Madre

La tabla 4.7 muestra el valor p de la prueba de independencia entre las variables, el valor p resultante es menor que 0.05, por lo tanto se concluye que estas variables son dependientes entre sí.

**Tabla 4.7 Tabla de contingencia. Variables Desertor vs. Formación de la Madre**

Nivel de Formación de la Madre	No Desertor	Desertor	Total
Más de secundaria	103 38.4%	18 17.5%	121 32.6%
Secundaria o menos	119 44.4%	15 14.6%	134 36.1%
No sabe	46 17.2%	70 68.0%	116 31.3%
Total	268 100.0%	103 100.0%	371 100.0%

Ji-cuadrado = 89.765

g.l. = 2

p= 0.000

Fuente: Secretaría Técnica Académica  
Elaboración: Autor



#### 4.2.9 Desertor vs. Reprobó Materias

La tabla 4.11 muestra el valor p de la prueba de independencia entre las variables, el valor p resultante es menor que 0.05, por lo tanto se concluye que estas variables son dependientes entre sí.

**Tabla 4.11 Tabla de contingencia. Variables Desertor vs. Reprobó Materias**

Reprobó Materias	No Desertor	Desertor	Total
No	134 50.0%	20 19.4%	154 41.5%
Sí	134 50.0%	83 80.6%	217 58.5%
Total	268 100.0%	103 100.0%	371 100.0%

Ji-cuadrado = 28.662      g.l. = 1      p= 0.000

Fuente: Secretaría Técnica Académica  
Elaboración: Autor

#### 4.2.10 Desertor vs Promedio General

Al realizar la prueba de Levene para igualdad de varianzas (tabla 4.12) se obtiene un valor p menor a 0.05 por lo que no es posible asumir que las varianzas son iguales; por esta razón suponemos que las varianzas no son iguales, con lo que al realizar la prueba t de student para la igualdad de medias se obtiene un valor p menor que 0.05 lo que indica que sí existen diferencias significativas en las medias de los promedios generales de los desertores vs. no desertores.

**Tabla 4.12 Prueba de Levene. Variable Promedio General**

	F	Sig.
<i>Promedio General</i>	57.308	0.000

Fuente: Secretaría Técnica Académica  
Elaboración: Autor

**Tabla 4.13 Prueba t de student. Variable Promedio General**

	T	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
						Inferior	Superior
Promedio General	6.713	113.828	0.000	1.13586	0.16921	0.80066	1.47107

Fuente: Secretaría Técnica Académica  
Elaboración: Autor

#### 4.2.11 Desertor vs Aprobó Cálculo

La tabla 4.14 muestra el valor p de la prueba de independencia entre las variables, el valor p resultante es menor que 0.05, por lo tanto se concluye que estas variables son dependientes entre sí.

**Tabla 4.14** Tabla de contingencia. Variables Desertor vs. Aprobó cálculo

Aprobó calculo	No Desertor	Desertor	Total
	155	24	179
Sí	57.8%	23.3%	48.2%
	89	68	157
No	33.2%	66.0%	42.3%
	24	11	35
No tomó materia	9.0%	10.7%	9.4%
	268	103	371
Total	100%	100%	100%

Ji-cuadrado = 37.554 g.l. = 2 p= 0.000

Fuente: Secretaría Técnica Académica  
Elaboración: Autor

#### 4.2.12 Desertor vs. Aprobó Física A

La tabla 4.15 muestra el valor p de la prueba de independencia entre las variables, el valor p resultante es menor que 0.05, por lo tanto se concluye que estas variables son dependientes entre sí.

**Tabla 4.15** Tabla de contingencia. Variables Desertor vs. Aprobó física

Aprobó Física A	No Desertor	Desertor	Total
	163	28	191
Sí	60.8%	27.2%	51.5%
	61	51	112
No	22.8%	49.5%	30.2%
	44	24	68
No tomó materia	16.4%	23.3%	18.3%
	268	103	371
Total	100.0%	100.0%	100.0%

Ji-cuadrado = 35.91 g.l. = 2 p= 0.000

Fuente: Secretaría Técnica Académica  
Elaboración: Autor

#### 4.2.13 Desertor vs Aprobó Química General

La tabla 4.16 muestra el valor p de la prueba de independencia entre las variables, el valor p resultante es menor que 0.05, por lo tanto se concluye que estas variables son dependientes entre sí.

**Tabla 4.16 Tabla de contingencia. Variables Desertor vs. Aprobó química**

Aprobó Química	No Desertor	Desertor	Total
Sí	225 84.0%	64 62.1%	289 77.9%
No	3 1.1%	17 16.5%	20 5.4%
No tomó materia	40 14.9%	22 21.4%	62 16.7%
Total	268 100.0%	103 100.0%	371 100.0%

Ji-cuadrado = 39.061      g.l. = 2      p= 0.000

Fuente: Secretaría Técnica Académica  
Elaboración: Autor

#### 4.3. Regresión Logística

La regresión logística tiene como objetivo modelar una variable dicotómica, que representa la aparición u ocurrencia de un determinado suceso, a través de un conjunto de variables predictoras cuantitativas o cualitativas.

Como se mencionó en capítulos previos, la variable a explicar con esta técnica será *Desertó*, la cual indica si un estudiante abandonó sus estudios por al menos dos términos consecutivos durante los primeros cinco años de su carrera.

Luego de explorar las relaciones bivalentes entre la variable dependiente y las variables predictoras, tomadas de una en una, se procedió a realizar modelos de regresión logística con cada una de las variables explicativas, de modo que se confirmó las relaciones identificadas previamente.

Además se consideró relevante incluir en el modelo de regresión la variable *Tipo de Colegio* puesto que al incluirla en la regresión logística, esta variable resultó estadísticamente significativa; no así la variable *Promedio General*, la cual a pesar de ser estadísticamente significativa se decidió no incluirla ya que podría generar

multicolinealidad en el modelo por la alta correlación que tiene con las variables *Aprobó Cálculo*, *Aprobó Física* y *Aprobó Química* originales (cuantitativas, sin recodificar). Es así que el modelo final se compuso de las siguientes variables:

**Tabla 4.17 Codificación de variables del modelo de regresión logística**

Variables		Codificación
Variable Dependiente	<i>Desertor</i>	0: No 1: Sí
	<i>Edad_Termino</i>	Variable cuantitativa
Variables Predictoras	<i>Sexo</i>	0: Hombre 1: Mujer
	<i>Porcentaje_crédito</i>	0: >=75 1: 50.01-75 2: 25.01-50 3: <=25
	<i>Formación_padre</i>	0: Más de secundaria 1: Secundaria o menos 2: No sabe
	<i>Formación_madre</i>	0: Más de secundaria 1: Secundaria o menos 2: No sabe
	<i>Miembros_Hogar</i>	Variable cuantitativa
	<i>Tipo_colegio</i>	0: Particular 1: Fiscomisional o Fiscal
	<i>Reprobó_materias</i>	0: No 1: Sí
	<i>Aprobó_Cálculo</i>	0: Sí 1: No 2: No tomó la materia
	<i>Aprobó_Física</i>	0: Sí 1: No 2: No tomó la materia
	<i>Aprobó_Química</i>	0: Sí 1: No 2: No tomó la materia

Fuente: Secretaría Técnica Académica  
Elaboración: Autor

#### 4.3.1 Resultados del modelo

El modelo de regresión logística estimado está compuesto por 18 variables independientes, en la tabla 4.18 se muestran los *B* estimados, error estándar, estadístico de wald y valor *p*.

**Tabla 4.18 Coeficientes B estimados del modelo Logit**

Grupo de Control	Covariable	B	Error Estándar	Estadístico de Wald	Valor p
-	Constante	-2.290	3.297	0.482	0.487
Sexo (Hombre)	Mujer	-0.395	0.401	0.970	0.325
	Edad	-0.041	0.175	0.056	0.813
Formación del Padre (Más de secundaria)	Secundaria o menos	1.621	0.587	7.635	0.006
	No sabe	2.063	0.571	13.061	0.000
Formación de la Madre (Más de secundaria)	Secundaria o menos	-0.270	0.494	0.299	0.585
	No sabe	1.632	0.495	10.893	0.001
Porcentaje a pagar por crédito (>=75)	50.01-75	-2.186	0.714	9.383	0.002
	25.01-50	-2.493	0.561	19.715	8.99e-06
	<= 25	-3.952	0.724	29.836	4.70e-08
Aprobó Cálculo Diferencial (Sí)	No	1.272	0.532	5.724	0.017
	No tomó materia	-1.480	0.903	2.684	0.101
Aprobó Física A (Sí)	No	0.942	0.462	4.150	0.042
	No tomó materia	0.704	1.051	0.449	0.503
Aprobó Química (Sí)	No	2.078	0.943	4.853	0.028
	No tomó materia	0.762	0.931	0.670	0.413
Reprobó Materias (No)	Sí	0.160	0.599	0.071	0.789
	Número de miembros del hogar	0.304	0.156	3.819	0.051
Tipo de Colegio (Particular)	Fiscal o Fiscomisional	0.947	0.445	4.532	0.033

Fuente: Secretaría Técnica Académica  
Elaboración: Autor

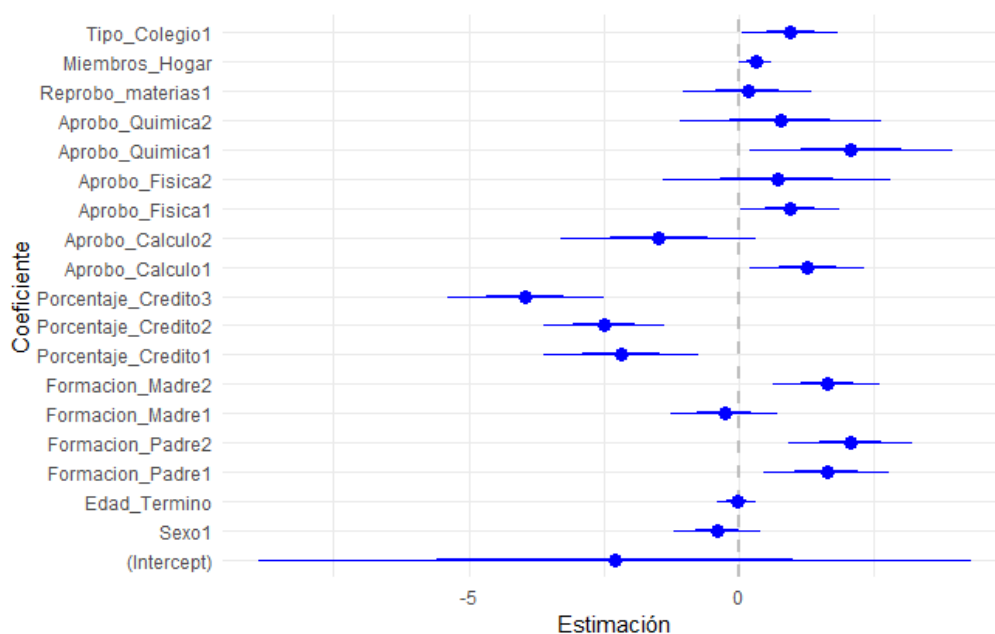
Al analizar los signos de los coeficientes de la tabla 4.18 encontramos siete variables con signos negativos, estas son: *Sexo - Mujer*, *Edad*, *Formación de la Madre - Secundaria o menos*, *Porcentaje a pagar por crédito 50.01-75*, *Porcentaje a pagar por crédito 25.01-50*, *Porcentaje a pagar por crédito menor o igual que 25* y *Cálculo – No tomó materia*; estas variables representan factores de protección de la deserción, mientras que todas las demás representan factores de riesgo.



Al contrastar las covariables con la variable dependiente *Desertor*, un poco más de la mitad de las covariables del modelo resultaron estadísticamente significativas (valor p menor que 0.05), con excepción de las siguientes variables: *Sexo*, *Edad*, *Formación de la Madre* – secundaria o menos, *Física A* – No tomó materia, *Química* – no tomó materia y *Reprobó materias* – Sí.

Los coeficientes estimados con sus respectivos intervalos de confianza se muestran en la gráfica 4.1.

**Figura 4.1 Coeficientes estimados del modelo**



Fuente: Secretaría Técnica Académica  
Elaboración: Autor

La oportunidad relativa (*Odds Ratio*), es la comparación de la probabilidad de ocurrencia o aparición de un evento luego de la exposición a un determinado factor de riesgo, frente a la probabilidad de ocurrencia del mismo evento en una situación o estado de control (Andrade, 2015). En la tabla 4.19 se muestran las oportunidades relativas del modelo estimado, su intervalo de confianza, y su interpretación.

**Tabla 4.19 Oportunidades relativas**

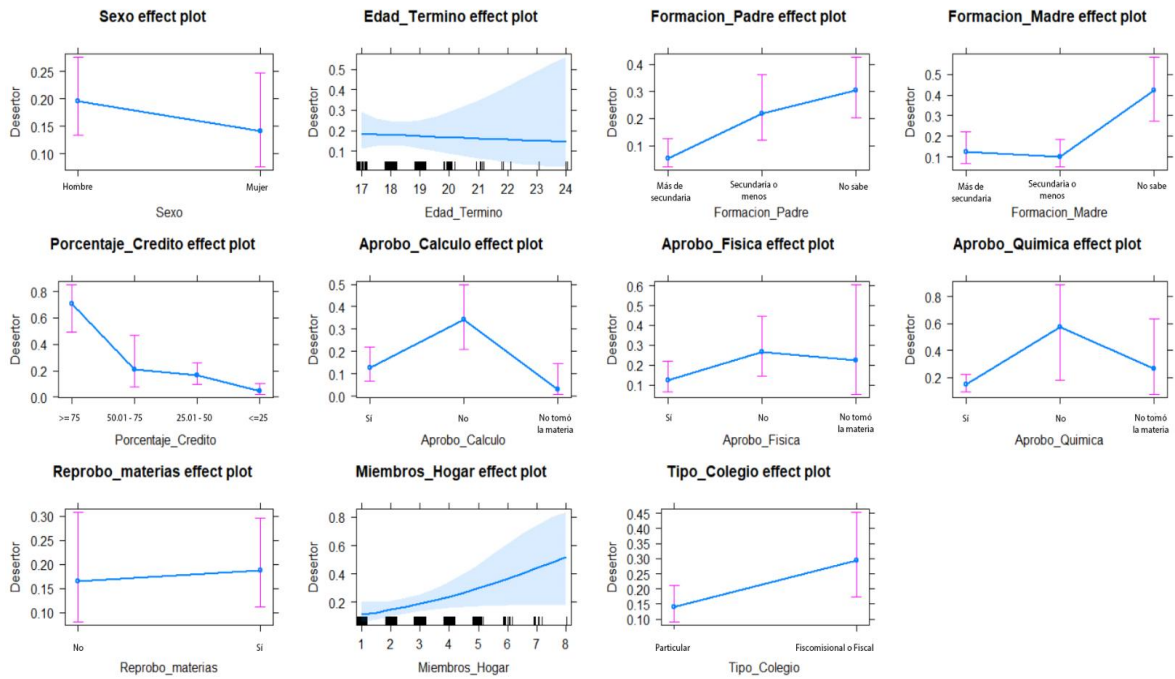
Grupo de Control	Covariable	Exp(B)	I.C. 95% para Exp(B)		Interpretación
			Inf.	Sup.	
Sexo (Hombre)	Mujer	0.674	0.307	1.479	Ser mujer disminuye en un 67% la probabilidad de desertar en relación a ser hombre
	Edad	0.959	0.680	1.353	La oportunidad de desertar disminuye en 0.959 por cada año cumplido
Nivel de Formación del Padre (Más de secundaria)	Secundaria o menos	5.058	1.602	15.973	La oportunidad de desertar es 5 veces más alta al tener el padre del estudiante nivel de formación secundaria o menos, en comparación con aquellos cuyos padres tienen formación superior
	No sabe	7.867	2.570	24.077	La oportunidad de desertar es 7 veces más alta en los estudiantes que desconocen el nivel de formación de su padre, en comparación con aquellos cuyos padres tienen formación superior
Nivel de Formación de la Madre (Más de secundaria)	Secundaria o menos	0.763	0.290	2.010	Al tener la madre nivel de formación secundaria o menor, disminuye en un 76% la oportunidad de desertar, en relación a tener madre con nivel de formación superior.
	No sabe	5.116	1.941	13.487	La oportunidad de desertar es 5 veces más alta en los estudiantes que desconocen el nivel de formación de su madre, en comparación con aquellos cuyas madres tienen formación superior.
Porcentaje a pagar por crédito (>=75)	50.01-75	0.112	0.028	0.455	Tener un porcentaje a pagar por crédito inferior al 75% disminuye la oportunidad de desertar.
	25.01-50	0.083	0.028	0.248	
	menor o igual que 25	0.019	0.005	0.079	
Aprobó Cálculo Diferencial (Sí)	No	3.568	1.259	10.116	Un estudiante que reprueba cálculo diferencial es 3 veces más propenso a desertar respecto a los que aprueban la materia.
	No tomó materia	0.228	0.039	1.337	No tomar la materia de cálculo diferencial disminuye en 0.228 la oportunidad de desertar, en relación a aquellos que aprueban la materia

Grupo de Control	Covariable	Exp(B)	I.C. 95% para Exp(B)		Interpretación
			Inf.	Sup.	
Aprobó Física A (Sí)	No	2.564	1.036	6.346	Un estudiante que reprueba Física A es 2 veces más propenso a desertar respecto a los que aprueban la materia
	No tomó materia	2.022	0.258	15.857	Un estudiante que no toma la materia Física A es 2 veces más propenso a desertar respecto a los que sí aprueban la materia
Aprobó Química (Sí)	No	7.986	1.258	50.713	Un estudiante que reprueba Química General es 7 veces más propenso a desertar respecto a los que sí aprueban la materia
	No tomó materia	2.143	0.346	13.291	Un estudiante que no toma la materia Química General es 2 veces más propenso a desertar respecto a los que sí aprueban la materia
Reprobó Materias (No)	Sí	1.174	0.363	3.799	Los estudiantes que reprobaron al menos una materia son 1.174 veces más propensos a desertar en relación a aquellos que no reprobaron materias
Número de miembros del hogar		1.356	0.999	1.839	La oportunidad de desertar aumenta en 1.356 por cada miembro en el hogar del estudiante.
Tipo de Colegio (Particular)	Fiscal o Fiscomisional	2.579	1.078	6.170	Los estudiantes provenientes de colegios fiscales o fiscomisionales son 2 veces más propensos a desertar en relación a aquellos que provienen de colegios particulares

Fuente: Secretaría Técnica Académica  
Elaboración: Autor

A continuación se muestran los efectos de cada uno de los factores sobre la probabilidad de desertar de forma gráfica, con lo cual es posible observar la magnitud del cambio de la variable *Desertor* a medida que cambia la variable independiente, mientras que el resto de las variables son controladas (ajustadas) considerando para ello los valores promedio o típicos. Los gráficos incluyen además barras de error que permiten tener una idea visual de la incertidumbre de las estimaciones.

**Figura 4.2 Efectos de los factores del modelo sobre la probabilidad de desertar**



Fuente: Secretaría Técnica Académica  
Elaboración: Autor

Con los coeficientes  $B$  estimados de la tabla 4.19 se construye la ecuación del modelo, donde  $y$  es  $P(y = 1)$  es la probabilidad de que un estudiante de carreras del área STEM deserte al menos una vez durante los primeros cinco años de su carrera universitaria.

$$P(y = 1|x) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

Donde

$$z = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p$$

$$\begin{aligned}
 (y = 1 | x) = & 1 / (1 - e^{(2.290 + 0.395 \text{ Sexo mujer} + 0.041 \text{ Edad} \\
 & - 1.621 \text{ Formación del Padre secundaria o menos} \\
 & - 2.063 \text{ Formación del Padre no sabe} \\
 & + 0.270 \text{ Formación de la Madre secundaria o menos} \\
 & - 1.632 \text{ Formación de la Madre no sabe} \\
 & + 2.186 \text{ Porcentaje a pagar por crédito entre 50.01 y 75} \\
 & + 2.493 \text{ Porcentaje a pagar por crédito entre 25.01 y 50} \\
 & + 3.952 \text{ Porcentaje a pagar por crédito menor o igual que 25} \\
 & - 1.272 \text{ Reprobó Cálculo} + 1.480 \text{ No tomó Cálculo} \\
 & - 0.942 \text{ Reprobó Física A} - 0.704 \text{ No tomó Física A} \\
 & - 2.078 \text{ Reprobó Química} - 0.762 \text{ No tomó Química} \\
 & - 0.160 \text{ Reprobó Materias} - 0.304 \text{ Número de miembros del Hogar} \\
 & - 0.947 \text{ Tipo de Colegio Fiscal o Fiscomisional} ) )
 \end{aligned}$$

### 4.3.2 Ajuste del modelo

Una vez estimado el modelo, procedemos a evaluar su bondad de ajuste, para esto utilizaremos varios métodos; el primero es el pseudo  $R^2$ , que mide qué tan bien predicen las variables independientes a la variable dependiente. El pseudo  $R^2$  de McFadden obtenido es de 0.49, valores entre 0.2 y 0.4 indican un buen ajuste del modelo.

Otra medida del ajuste es la significancia global del modelo. Esta prueba evalúa si el modelo con predictores se ajusta significativamente mejor que un modelo con solo un intercepto (es decir, el modelo nulo). El estadístico de prueba es la diferencia entre la desviación residual del modelo con las variables predictoras y la desviación residual del modelo nulo, sigue una distribución Ji-cuadrado con grados de libertad iguales a las diferencias en grados de libertad entre el modelo actual y nulo (es decir, el número de variables predictoras el modelo). Para nuestro modelo el valor p obtenido es muy cercano a 0, lo que indica que nuestro modelo se ajusta significativamente mejor que el modelo nulo.

Por otro lado, la prueba de Hosmer y Lemeshow compara por intervalos los valores observados con los esperados, sigue una distribución Ji-cuadrado y su hipótesis nula manifiesta que no existen diferencias entre los valores observados y los esperados; para nuestro modelo el valor p obtenido es de 0.4178, por lo que no existe evidencia estadística suficiente para rechazar la hipótesis nula, lo que nos permite concluir que modelo se ajusta bien a los datos.

**Tabla 4.20 Medidas de bondad de ajuste del modelo**

Pseudo $R^2$ McFadden	Significación Global			Prueba de Hosmer - Lemeshow		
	$\chi^2$	gl.	Sig.	H - L	gl.	Sig.
0.49	216.26	18	5.48e-24	17.549	17	0.4178

*Fuente: Secretaría Técnica Académica  
Elaboración: Autor*

Además es posible evaluar la capacidad predictiva del modelo a través de la construcción de la tabla de confusión, en esta tabla se muestran los valores observados de la variable *Desertó*, distinguiendo entre desertores y no desertores, así como los valores estimados por el modelo para los cuales se establece un punto de corte para las probabilidades, comúnmente en 0.5 para clasificar a los estudiantes de la muestra en uno de los dos grupos.

El modelo estimado tuvo un porcentaje global del 88.7% de aciertos, de los cuales se encuentran clasificados correctamente el 72.8% de los estudiantes desertores, y el 94.9% de los no desertores; estos resultados indican que la capacidad predictiva del modelo es alta.

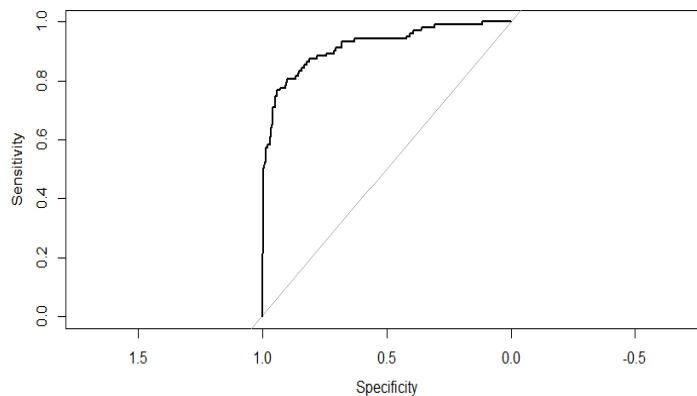
**Tabla 4.21 Tabla de clasificación**

Observados	Pronosticados		
	No	Sí	Porcentaje correcto
No	254	14	94.8
Sí	28	75	72.8
Porcentaje global			88.7

*Fuente: Secretaría Técnica Académica  
Elaboración: Autor*

Finalmente la curva ROC muestra gráficamente la sensibilidad y especificidad del modelo, el área bajo de la curva se denomina AUC, y toma valores entre 0 y 1, mientras más próximo sea el valor a 1 significa que el modelo permite predecir de mejor manera los dos posibles estados de la variable de respuesta. El AUC del modelo estimado es de 0.9168 con lo que se confirma una vez más que el modelo se ajusta bien a los datos.

**Figura 4.3 Curva ROC**



*Fuente: Secretaría Técnica Académica  
Elaboración: Autor*

## **CAPÍTULO 5**

### **5. CONCLUSIONES Y RECOMENDACIONES**

En este trabajo se estimó un modelo para determinar la probabilidad de desertar al menos una vez durante los primeros años de estudio, de estudiantes de carreras del área STEM en una institución de educación superior pública, con base en datos registrados en la Secretaría Técnica Académica de dicha institución.

A través de las técnicas y análisis empleados hemos llegado a las siguientes conclusiones y recomendaciones:

#### **5.1 CONCLUSIONES**

- Durante el análisis bivariado se encontró que la edad al ingreso a la carrera, el sexo, el porcentaje a pagar por crédito, el nivel de formación de los padres, el número de miembros del hogar, el reprobar materias, el promedio general y la aprobación de las materias de Cálculo Diferencial, Física A y Química General son variables que se asocian significativamente con el abandono de la universidad.
- El modelo de regresión logística permitió conocer los efectos de las variables utilizadas para predecir la deserción. El mismo permite inferir que el comenzar la carrera a una edad mayor, el bajo nivel de escolaridad de los padres, un mayor número de miembros en el hogar y reprobar materias núcleo del ciclo básico común de ingenierías tales como Cálculo Diferencial, Física A y Química General, se asocian con una mayor propensión a desertar de la universidad.
- Los factores que se identificaron como de protección de la deserción son ser mujer, tener madre con baja escolaridad y, contrario a lo que podría esperarse, el tener un porcentaje a pagar por crédito inferior al 75%.



- El modelo estimado permite determinar la probabilidad de deserción de un estudiante de carreras del área STEM; con una precisión del 88.7% y de acuerdo a las pruebas de bondad de ajuste realizadas se ajusta bien a los datos.
- El análisis, evaluación y aplicación de los resultados obtenidos en el presente trabajo permitirá soportar la toma de decisiones, así como orientar la formulación de políticas, lineamientos y estrategias enfocados en reducir la deserción estudiantil universitaria.

## **5.2 RECOMENDACIONES**

- Entre las dificultades presentadas para realizar esta investigación, la más influyente fue la mala calidad de los datos almacenados en la institución, lo que incurrió en que aún después de la depuración de los datos, se descartaran variables que podrían haber resultado relevantes en este estudio. Por lo descrito se recomienda mejorar los procesos de recolección y sistematización de la información, así como también integrar las bases de datos institucionales para proveer reportes consolidados y útiles para la toma de decisiones. Otra acción sugerida es la actualización periódica la información de los estudiantes de las instituciones de educación superior de modo que exista disponibilidad de datos relevantes sobre el estudiante, en especial sobre variables socioeconómicas y académicas que permitan realizar estudios de mayor profundidad. Es necesario evaluar si los datos que actualmente dispone la IES son los suficientes, y si son lo suficientemente confiables para tomar decisiones a partir de ellos.
- Implementar un sistema de alertas tempranas, a partir del modelo estimado, de modo que permita a las áreas encargadas diseñar estrategias para mitigar la deserción estudiantil; y formular políticas y programas orientados a ayudar a aquellos estudiantes que de acuerdo a los análisis realizados son más propensos a abandonar sus estudios universitarios. Como ejemplo, aquellos estudiantes que el sistema alerte como posibles desertores podrían

recibir un programa de tutorías especial, que brinde apoyo no solo académico, sino también motivacional.

- Estandarizar los requerimientos de información a las instituciones de educación superior, a nivel de país, y poner a disposición de la ciudadanía las bases de datos consolidadas, de modo que permitan realizar estudios a nivel nacional y comparar los resultados obtenidos en este trabajo con otros similares en otro tipo de instituciones.

## 6. REFERENCIAS

- [1]. Calderón Defaz , M., Espinel Armas, E., Garzón Carrera, P., & Pástor Guevara, C. (2017). Impacto social de la deserción estudiantil en la Facultad de Ciencias Químicas en primer semestre de la Universidad Central del Ecuador. *Polo del Conocimiento*, 2(8), 65-89.
- [2]. Andrade, C. (2015). Understanding relative risk, odds ratio, and related terms: as simple as it can get. *The Journal of clinical psychiatry*, 857-861.
- [3]. Canavos, G. C. (1988). *Probabilidad y Estadística. Aplicaciones y Métodos*. México: Mcgraw-hill.
- [4]. CINDA. (2006). *Repitencia y deserción universitaria en América Latina*. Santiago: Alfabeta artes gráficas.
- [5]. Everitt, B., & Dunn, G. (2001). *Applied Multivariate Data Analysis*. London: Arnold.
- [6]. Hosmer, D., & Lemeshow, S. (2013). *Applied logistic regression*. John Wiley & Sons.
- [7]. Marta, M., Avitabile, C., Botero, J., Haimovich, F., & Urzúa, S. (2017). *Momento decisivo; la educación superior en América Latina y el Caribe*. Washington, DC. Obtenido de <https://openknowledge.worldbank.org/handle/10986/26489>
- [8]. NU. CEPAL, CAF, & OCDE. (2016). *Perspectivas económicas de América Latina 2017: Juventud, Competencias y Emprendimiento*. París: OECD Publishing.
- [9]. Rama, C. (2009). La tendencia a la masificación de la cobertura de la educación superior en América Latina. *Revista Interamericana de la Educación/Revista(50)*, 173-195.
- [10]. Rama, C. (2016). La educación superior en América Latina ante los nuevos desafíos del ciclo económico. En R. R. G., *UNIVERSIDAD URGENTE para una sociedad emancipada* (págs. 73-80). Quito: SENESCYT-IESALC.

- [11]. Ramírez, R. (2016). Universidad urgente para una sociedad emancipada. En R. Ramírez, *UNIVERSIDAD URGENTE para una sociedad emancipada* (pág. 30). Quito: SENESCYT-IESALC.
- [12]. Rosado Gómez, A. A. (2013). Modelo de detección de estudiantes excluidos en carreras de ingeniería utilizando Minería de Datos. *Revista Ingenio UFPSO*.
- [13]. Roth, S. E. (2008). A model to predict Ohio University student attrition from admissions and involvement data. *Tesis doctoral. Ohio University*.
- [14]. Rubio Hurtado, M. J., & Berlanga Silvente, V. (2012). Cómo aplicar las pruebas paramétricas bivariadas t de Student y ANOVA en SPSS. *REIRE. Revista d'Innovació i Recerca en Educació*, 5, p. 83-100.
- [15]. Seldaña Villa, M., & Barriga, O. (2010). Adaptación del modelo de deserción universitaria de Tinto a la Universidad Católica de la Santísima Concepción, Chile. *Revista de Ciencias Sociales*, 16, 616-628.
- [16]. SENESCYT. (2016). *Rendición de Cuentas: Secretaría de Educación Superior, Ciencias, Tecnología e Innovación*. Recuperado el 8 de Agosto de 2018, de Educación Superior: <https://www.educacionsuperior.gob.ec/>
- [17]. SENESCYT. (2017). *Educación Superior*. Recuperado el 25 de Septiembre de 2018, de Rendición de Cuentas: Secretaría de Educación Superior, Ciencias, Tecnología e Innovación: <http://www.senescyt.gob.ec/rendicion>
- [18]. Smith, T. J., & McKenna, C. M. (2013). A comparison of logistic regression pseudo R2 indices. *Multiple Linear Regression Viewpoints*, 39(2), 17-26.
- [19]. Trevizán, A. L., Beltrán, C., & Cosolito, P. (2009). Variables que condicionan la deserción y retención durante el trayecto universitario de alumnos de carrera de Ingeniería Agronómica de la Universidad Nacional de Rosario. *Revista de epistemología y ciencias humanas*, 1, 85-95.
- [20]. Zurita Herrera, G. (2008). *Probabilidad y Estadística, Fundamentos y Aplicaciones*. Guayaquil: ICM - ESPOL.

## 7. ANEXOS

### Anexo A. Código de R

```
> data <- read.delim(file.choose(), header = TRUE, sep="\t")
> attach(data)
> str(data)
'data.frame': 371 obs. of 36 variables:
 $ DESERTOR : int 0 0 0 0 0 0 0 0 0 0 ...
 $ APROBAR_CALCULO : int 2 2 2 2 2 2 2 2 2 2 ...
 $ APROBAR_FISICA : int 2 2 2 2 2 2 2 2 2 2 ...
 $ APROBAR_QUIMICA : int 2 2 2 2 2 2 2 2 2 2 ...
 $ MREP : int 0 0 0 0 0 0 0 0 0 0 ...
 $ EDAD_TERMINO : int 18 18 18 20 18 19 18 18 21 19 ...
 $ SEXO : int 0 0 1 0 0 0 1 0 0 0 ...
 $ ESTADOCIVIL_RECO : int 0 0 0 0 0 0 0 0 0 0 ...
 $ FORMACIONPADRE_RECO: int 2 0 0 1 2 2 2 1 2 1 ...
 $ FORMACIONMADRE_RECO: int 0 0 1 2 2 0 2 1 2 1 ...
 $ TIPOCOLEGIO_RECO : int 0 0 0 0 0 0 0 0 0 1 ...
 $ MIEMBROS_HOGAR : int 2 2 3 2 1 3 1 1 3 3 ...
 $ PORCENTAJE_CREDITO : int 1 2 2 3 0 2 0 0 1 3 ...
 $ RECIBE_BONO : int 1 1 1 1 1 1 1 1 1 1 ...
 $ PORCENTAJE_A_PAGAR : Factor w/ 207 levels "10","10,2","10,5",...: 165
151 123 53 5 127 5 207 181 4 ...
 $ MATERIAS_REPROBADAS: int 0 0 0 0 0 0 0 0 0 0 ...
 $ MATERIAS_APROBADAS : int 5 7 6 6 7 6 6 7 6 5 ...
 $ PROMEDIO_GENERAL : Factor w/ 208 levels ",35","0","1,58",...: 177 91
193 136 172 191 201 161 163 195 ...
 $ UNIDAD_ACADEMICA : int 3 5 5 1 2 1 5 2 1 1 ...
 $ CALCULO : Factor w/ 113 levels " ",",3","8",...: 1 1 1 1 1
1 1 1 1 1 ...
 $ FISICAA : Factor w/ 96 levels " ",",35","0",...: 1 1 1 1 1
1 1 1 1 1 ...
 $ QUIMICA : Factor w/ 90 levels " ",",3","0","10",...: 1 1 1
1 1 1 1 1 1 ...
 $ INGRESOS3 : int 0 0 NA 1 NA 0 0 0 0 1 ...
 $ ETNIA_RECO : int 0 0 0 0 0 0 0 0 0 1 ...
 $ ESTRATO_RECO : int NA 0 0 1 NA 0 NA NA NA 1 ...
 $ DONDERESIDE_RECO : int NA 1 0 0 NA 0 NA NA NA 0 ...
 $ TIPO_VIVIENDA : int NA 2 0 0 NA 0 NA NA NA 2 ...
 $ BECA_COLEGIO : int NA 1 1 1 NA 1 NA NA NA 1 ...
 $ VIVIENDA_PROPIA : int NA 1 0 0 NA 0 NA NA NA 1 ...
 $ PC : int 0 0 1 1 1 1 0 0 NA 1 ...
 $ LAPTOP : int 0 1 0 0 0 0 1 0 NA 0 ...
 $ CELULAR : int 0 0 0 0 0 0 0 0 NA 0 ...
 $ PLAN_CELULAR : int 1 1 0 1 1 1 0 1 NA 1 ...
 $ MATERIAS_CURSANDO : int 5 7 6 6 7 6 6 7 6 5 ...
 $ GRADUADO : int 0 1 0 1 0 0 0 0 1 0 ...
 $ COD_ESTUDIANTE : int 201102982 201112812 201112910 201112950 201
116682 201135707 201144257 201152216 201154841 201155067 ...
> colnames(data)[1] <- "Desertor"
> colnames(data)[2] <- "Aprobo_Calculo"
> colnames(data)[3] <- "Aprobo_Fisica"
> colnames(data)[4] <- "Aprobo_Quimica"
> colnames(data)[5] <- "Reprobo_materias"
> colnames(data)[6] <- "Edad_Termino"
> colnames(data)[7] <- "Sexo"
> colnames(data)[9] <- "Formacion_Padre"
> colnames(data)[10] <- "Formacion_Madre"
> colnames(data)[11] <- "Tipo_Colegio"
> colnames(data)[12] <- "Miembros_Hogar"
> colnames(data)[13] <- "Porcentaje_Credito"
> data$Aprobo_Calculo<-as.factor(data$Aprobo_Calculo)
> data$Aprobo_Fisica<-as.factor(data$Aprobo_Fisica)
> data$Aprobo_Quimica<-as.factor(data$Aprobo_Quimica)
> data$Reprobo_materias<-as.factor(data$Reprobo_materias)
> data$Sexo<-as.factor(data$Sexo)
> data$ESTADOCIVIL_RECO<-as.factor(data$ESTADOCIVIL_RECO)
```

```

> data$Formacion_Madre<-as.factor(data$Formacion_Madre)
> data$Formacion_Padre<-as.factor(data$Formacion_Padre)
> data$Tipo_Colegio<-as.factor(data$Tipo_Colegio)
> data$Porcentaje_Credito<-as.factor(data$Porcentaje_Credito)
>
> data$RECIBE_BONO<-as.factor(data$RECIBE_BONO)
> data$UNIDAD_ACADEMICA<-as.factor(data$UNIDAD_ACADEMICA)
> data$INGRESOS3<-as.factor(data$INGRESOS3)
> data$ETNIA_RECO<-as.factor(data$ETNIA_RECO)
> data$ESTRATO_RECO<-as.factor(data$ESTRATO_RECO)
> data$DONDERESIDE_RECO<-as.factor(data$DONDERESIDE_RECO)
> data$TIPO_VIVIENDA<-as.factor(data$TIPO_VIVIENDA)
> data$BECA_COLEGIO<-as.factor(data$BECA_COLEGIO)
> data$VIVIENDA_PROPIA<-as.factor(data$VIVIENDA_PROPIA)
> data$PC<-as.factor(data$PC)
> data$LAPTOP<-as.factor(data$LAPTOP)
> data$CELULAR<-as.factor(data$CELULAR)
> data$PLAN_CELULAR<-as.factor(data$PLAN_CELULAR)
> data$GRADUADO<-as.factor(data$GRADUADO)
> data$PROMEDIO_GENERAL<-as.numeric(data$PROMEDIO_GENERAL)
> data$PORCENTAJE_A_PAGAR<-as.numeric(data$PORCENTAJE_A_PAGAR)
> data$QUIMICA<-as.numeric(data$QUIMICA)
> data$CALCULO<-as.numeric(data$CALCULO)
> data$FISICAA<-as.numeric(data$PROMEDIO_GENERAL)
> attach(data)

```

The following objects are masked from data (pos = 3):

```

BECA_COLEGIO, CALCULO, CELULAR, COD_ESTUDIANTE, DONDERESIDE_RECO,
ESTADOCIVIL_RECO, ESTRATO_RECO, ETNIA_RECO, FISICAA, GRADUADO,
INGRESOS3, LAPTOP, MATERIAS_APROBADAS, MATERIAS_CURSANDO,
MATERIAS_REPROBADAS, PC, PLAN_CELULAR, PORCENTAJE_A_PAGAR,
PROMEDIO_GENERAL, QUIMICA, RECIBE_BONO, TIPO_VIVIENDA,
UNIDAD_ACADEMICA, VIVIENDA_PROPIA

```

```

> library("MASS")
> install.packages("ResourceSelection")
Installing package into 'C:/Users/sweet/OneDrive/Documents/R/win-library
/3.5'
(as 'lib' is unspecified)
also installing the dependency 'pbapply'

```

```

trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/pbapply_1.3
-4.zip'
Content type 'application/zip' length 67193 bytes (65 KB)
downloaded 65 KB

```

```

trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/ResourceSel
ection_0.3-2.zip'
Content type 'application/zip' length 498994 bytes (487 KB)
downloaded 487 KB

```

```

package 'pbapply' successfully unpacked and MD5 sums checked
package 'ResourceSelection' successfully unpacked and MD5 sums checked

```

```

The downloaded binary packages are in
C:\Users\sweet\AppData\Local\Temp\RtmpGUVU09\downloaded_packages

```

```

> library(ResourceSelection)
ResourceSelection 0.3-2      2017-02-28

```

```

> modelo<-glm(Desertor~Sexo+Edad_Termino+Formacion_Padre+Formacion_Madre
+Porcentaje_Credito+Aprobo_Calculo+Aprobo_Fisica+Aprobo_Quimica+Reprobo_
materias+Miembros_Hogar+Tipo_Colegio,family="binomial", data=data)
> summary(modelo)

```

```

Call:
glm(formula = Desertor ~ Sexo + Edad_Termino + Formacion_Padre +
  Formacion_Madre + Porcentaje_Credito + Aprobo_Calculo + Aprobo_Fisic
a +
  Aprobo_Quimica + Reprobo_materias + Miembros_Hogar + Tipo_Colegio,

```

```

family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5968 -0.4807 -0.2551  0.1053  2.9892

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.28957    3.29720  -0.694 0.487432
Sexo1       -0.39505    0.40120  -0.985 0.324783
Edad_Termino -0.04148    0.17538  -0.237 0.813023
Formacion_Padre1  1.62105    0.58666   2.763 0.005724 **
Formacion_Padre2  2.06263    0.57074   3.614 0.000302 ***
Formacion_Madre1 -0.27001    0.49402  -0.547 0.584685
Formacion_Madre2  1.63237    0.49459   3.300 0.000965 ***
Porcentaje_Credito1 -2.18597    0.71364  -3.063 0.002190 **
Porcentaje_Credito2 -2.49274    0.56140  -4.440 8.99e-06 ***
Porcentaje_Credito3 -3.95229    0.72356  -5.462 4.70e-08 ***
Aprobo_Calculo1  1.27201    0.53169   2.392 0.016738 *
Aprobo_Calculo2 -1.48001    0.90339  -1.638 0.101364
Aprobo_Fisica1  0.94171    0.46228   2.037 0.041641 *
Aprobo_Fisica2  0.70389    1.05091   0.670 0.502992
Aprobo_Quimica1  2.07766    0.94314   2.203 0.027601 *
Aprobo_Quimica2  0.76232    0.93103   0.819 0.412906
Reprobo_materias1 0.16022    0.59922   0.267 0.789171
Miembros_Hogar  0.30418    0.15565   1.954 0.050668 .
Tipo_Colegio1    0.94742    0.44505   2.129 0.033273 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

(Dispersion parameter for binomial family taken to be 1)

```

Null deviance: 438.30 on 370 degrees of freedom
Residual deviance: 222.04 on 352 degrees of freedom
AIC: 260.04

```

Number of Fisher Scoring iterations: 6

```

> pronos=predict(modelo,type = "response")
> pronitox=floor(pronos+0.5)
> table(pronitox,data$Desertor)

```

```

pronitox  0  1
          0 254 28
          1  14 75

```

```

> 254+75
[1] 329
> 329/371
[1] 0.8867925
> t1=hoslem.test(data$Desertor,pronos,19)
> t1

```

Hosmer and Lemeshow goodness of fit (GOF) test

```

data: data$Desertor, pronos
X-squared = 17.549, df = 17, p-value = 0.4178

```

```

> plot(sort(pronos,decreasing = TRUE))
> install.packages("pROC")
Installing package into 'C:/Users/sweet/OneDrive/Documents/R/win-library/3.5'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/pROC_1.12.1.zip'
Content type 'application/zip' length 1143312 bytes (1.1 MB)
downloaded 1.1 MB

```

package 'pROC' successfully unpacked and MD5 sums checked

The downloaded binary packages are in  
C:\Users\sweet\AppData\Local\Temp\RtmpGUVU09\downloaded\_packages

```
> library("pROC")  
Type 'citation("pROC")' for a citation.
```

Attaching package: 'pROC'

The following objects are masked from 'package:stats':

cov, smooth, var

```
> g<-roc(data$Desertor~pronos)  
> plot(g)  
> auc(g)  
Area under the curve: 0.9168  
> install.packages("pscl")  
Installing package into 'C:/Users/sweet/OneDrive/Documents/R/win-library  
/3.5'  
(as 'lib' is unspecified)  
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/pscl_1.5.2.  
zip'  
Content type 'application/zip' length 3536828 bytes (3.4 MB)  
downloaded 3.4 MB
```

package 'pscl' successfully unpacked and MD5 sums checked

The downloaded binary packages are in  
C:\Users\sweet\AppData\Local\Temp\RtmpGUVU09\downloaded\_packages

```
> library(pscl)  
Classes and Methods for R developed in the  
Political Science Computational Laboratory  
Department of Political Science  
Stanford University  
Simon Jackman  
hurdle and zeroinfl functions by Achim Zeileis
```

```
> pr2(modelo)  
          11h      11hNull          G2      McFadden          r2ML  
r2CU  
-111.0205398 -219.1493687 216.2576577 0.4934024 0.4417256 0.63  
72722
```

```
> with(modelo, df.null - df.residual)
```

```
[1] 18
```

```
> tg<-pchisq((438.3-222.04),18,lower.tail = F)
```

```
> tg
```

```
[1] 5.480619e-36
```

```
> exp(coef(modelo))
```

(Intercept)	Sexo1	Edad_Termino
0.10130973	0.67364637	0.95936752
Formacion_Padre1	Formacion_Padre2	Formacion_Madre1
5.05839042	7.86662017	0.76337287
Formacion_Madre2	Porcentaje_Credito1	Porcentaje_Credito2
5.11596143	0.11236907	0.08268278
Porcentaje_Credito3	Aprobo_Calculo1	Aprobo_Calculo2
0.01921071	3.56802860	0.22763590
Aprobo_Fisical	Aprobo_Fisica2	Aprobo_Quimical
2.56436245	2.02159469	7.98576189
Aprobo_Quimica2	Reprobo_materias1	Miembros_Hogar
2.14324019	1.17377378	1.35551159
Tipo_Colegio1		
2.57903942		

```
> exp(cbind(OR = coef(modelo), confint(modelo, level=0.95)))
```

Waiting for profiling to be done...

	OR	2.5 %	97.5 %
(Intercept)	0.10130973	0.0001159361	54.74720033
Sexo1	0.67364637	0.2998709246	1.45834977
Edad_Termino	0.95936752	0.6849997261	1.37205768
Formacion_Padre1	5.05839042	1.6681888296	17.07584959
Formacion_Padre2	7.86662017	2.7165193107	26.03467033



```

Formacion_Madre1    0.76337287  0.2858854016  2.00598087
Formacion_Madre2    5.11596143  1.9623783609 13.78930044
Porcentaje_Credito1 0.11236907  0.0249151284  0.42756240
Porcentaje_Credito2 0.08268278  0.0263642938  0.24112385
Porcentaje_Credito3 0.01921071  0.0042938986  0.07418162
Aprobo_Calculo1     3.56802860  1.3093495753 10.65801870
Aprobo_Calculo2     0.22763590  0.0370665556  1.29787261
Aprobo_Fisica1      2.56436245  1.0473841588  6.47284554
Aprobo_Fisica2      2.02159469  0.2667351270 16.22908541
Aprobo_Quimica1     7.98576189  1.4098769857 60.42416346
Aprobo_Quimica2     2.14324019  0.3356784837 12.82610732
Reprobo_materias1   1.17377378  0.3564122991  3.77868780
Miembros_Hogar      1.35551159  1.0004397248  1.84881550
Tipo_Colegio1       2.57903942  1.0907091012  6.29615502

```

```

> install.packages("coefplot")
Installing package into 'C:/Users/sweet/OneDrive/Documents/R/win-library/3.5'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/coefplot_1.2.6.zip'
Content type 'application/zip' length 139144 bytes (135 KB)
downloaded 135 KB

```

package 'coefplot' successfully unpacked and MD5 sums checked

The downloaded binary packages are in  
C:\Users\sweet\AppData\Local\Temp\RtmpGUVU09\downloaded\_packages

```

> library(coefplot)
Loading required package: ggplot2
> coefplot(modelo) +
+   theme_minimal() +
+   labs(title="Coeficientes estimados del modelo",
+         x="Estimación",
+         y="Coeficiente")

```

```

> install.packages("effects")
Installing package into 'C:/Users/sweet/OneDrive/Documents/R/win-library/3.5'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/effects_4.0-3.zip'
Content type 'application/zip' length 2259646 bytes (2.2 MB)
downloaded 2.2 MB

```

package 'effects' successfully unpacked and MD5 sums checked

The downloaded binary packages are in  
C:\Users\sweet\AppData\Local\Temp\RtmpGUVU09\downloaded\_packages

```

> library(effects)
Loading required package: carData
lattice theme set by effectsTheme()
see ?effectsTheme for details.
> allEffects(modelo)
model: Desertor ~ Sexo + Edad_Termino + Formacion_Padre + Formacion_Madre +
+   Porcentaje_Credito + Aprobo_Calculo + Aprobo_Fisica + Aprobo_Quimica
+   Reprobo_materias + Miembros_Hogar + Tipo_Colegio

```

Sexo effect  
Sexo

```

      0      1
0.1963357 0.1413156

```

Edad\_Termino effect  
Edad\_Termino

```

      17      19      20      22      24
0.1854876 0.1732791 0.1674171 0.1561699 0.1455462

```

Formacion\_Padre effect

```

Formacion_Padre
  0          1          2
0.05240267 0.21858630 0.30314975

Formacion_Madre effect
Formacion_Madre
  0          1          2
0.12490202 0.09825063 0.42203112

Porcentaje_Credito effect
Porcentaje_Credito
  0          1          2          3
0.70529125 0.21192793 0.16518823 0.04395392

Aprobo_Calculo effect
Aprobo_Calculo
  0          1          2
0.12646215 0.34060592 0.03190349

Aprobo_Fisica effect
Aprobo_Fisica
  0          1          2
0.1248541 0.2678546 0.2238520

Aprobo_Quimica effect
Aprobo_Quimica
  0          1          2
0.1451275 0.5754984 0.2667802

Reprobo_materias effect
Reprobo_materias
  0          1
0.1641517 0.1873332

Miembros_Hogar effect
Miembros_Hogar
  1          3          4          6          8
0.1121955 0.1884443 0.2394001 0.3664182 0.5151814

Tipo_Colegio effect
Tipo_Colegio
  0          1
0.1391333 0.2941963
> plot(allEffects(modelo))

```