ESPOCH Congresses: The Ecuadorian Journal of S.T.E.A.M.
Volume 3, Issue no. 1, DOI 10.18502/espoch.v3i1.14477
Production and Hosting by Knowledge E

Conference Paper

Predictive Model of Student Dropout Based on Logistic Regression

Modelo predictivo de deserción estudiantil basado en regresión logística

B.R. Cuji Chacha*, W.L. Gavilanes López, M.B. Pérez Constante

Universidad Técnica de Ambato (UTA), Ambato, Ecuador, 180110

ORCID
B.R. Cuji Chacha: https://orcid.org/0000-0003-4091-6876

IX CONGRESO INTERNACIONAL DE INVESTIGACIÓN DE LA RED ECUATORIANA DE UNIVERSIDADES Y ESCUELAS POLITÉCNICAS Y IX CONGRESO INTERNACIONAL DE CIENCIA TECNOLOGÍA EMPRENDIMIENTO E INNOVACIÓN SECTEI-ESPOCH 2022

Corresponding Author: B.R. Cuji Chacha; email: blancarcujic@uta.edu.ec

Published: 9 November 2023

Production and Hosting by Knowledge E

Abstract

Student desertion is a phenomenon that has spread significantly in many higher education institutions in Ecuador. The objective of the research was to develop a predictive model of student dropout based on multiple binary logistic regression, with the purpose of detecting possible dropouts. The methodology used consists of three phases: Phase 1: Analysis of variables; Phase 2: Formulation of the mathematical model; and Phase 3: Evaluation. For the estimation of the coefficients of the model, the SPSS tool was obtained. After the creation of the predictive model, it was concluded that the most significant variables that contribute to the diagnosis of dropout are marital status, age, gender, Note2s, and Note1s. It is also evident that students have a higher risk of dropping out if they are married and lower risk if they are single or divorced. Finally it was concluded that gender is a factor that directly influences dropout; male students are more likely to drop out than females.

Keywords: *logistic regression, predictive model, desertion.*

Resumen

La deserción estudiantil es un fenómeno que se ha extendido significativamente en gran cantidad de instituciones educativas de nivel superior en el Ecuador. El objetivo de la investigación fue desarrollar un modelo predictivo de deserción estudiantil basado en la regresión logística binaria múltiple, con el propósito de detectar a posibles desertores. La metodología utilizada consta de tres fases: Fase1: Análisis de variables. Fase2: Formulación del modelo matemático. Fase3: Evaluación. Para la estimación de los coeficientes del modelo se utilizó la herramienta SPSS. Posterior a la creación del modelo predictivo se llegó a concluir que las variables más significativas que aportan al diagnóstico de la deserción son estado civil, edad, género Nota2s y Nota1s, además se evidencia que los estudiantes tienen mayor riesgo de deserción si están casados y menor riesgo si están solteros o divorciados, finalmente se concluye, que el género es un factor que influye directamente en la deserción, los estudiantes masculinos son más propensos a desertar que los femeninos.

Palabras Clave: *regresión logística, modelo predictivo, deserción.*

OPEN ACCESS

1. Introducción

Actualmente, la deserción estudiantil se presenta como un problema del sistema de educación superior (1), por lo que es necesario afrontar diversas causas que están estrechamente ligadas al fenómeno. Factores socioeconómicos, curriculares, rendimiento académico, repetición, conducta, forman parte de la problemática (2), es importante enfrentarla proponiendo estrategias que optimicen los procesos educativos, determinando rasgos de estudiantes que sean proclives a desertar (3). Dentro del sistema educativo superior la deserción es tomada como un indicador que mide la eficiencia académica (4).

La deserción se produce cuando un estudiante inscrito en un determinado programa académico, no tiene continuidad regular durante el desarrollo del mismo, sea por causas relacionadas con retiros temporales o permanentes de la institución, o por repetir niveles fuera de su cohorte original (5). La deserción estudiantil o abandono escolar puede estar supeditada a la no matriculación de un determinado estudiante en un nivel superior (6). El fenómeno de la deserción estudiantil se percibe como la renuncia definitiva de la instrucción académica, por parte del estudiante por circunstancias internas, propias que afecta al individuo o externas relacionadas con el proceso enseñanza aprendizaje, las posibles causas asociadas a la deserción vienen dadas según la influencia que ejercen sobre el fenómeno, pudiendo ser clasificadas como poco o muy significativas (7). Además, es un problema que llega a representar elevados costos para la colectividad, varios factores inciden en este fenómeno, como la fecha de matriculación, asistencia a clases, nota de ingreso a la universidad, rendimiento académico principalmente en los primeros niveles de la carrera, siendo necesario que los estudiantes tengan una orientación vocacional apropiada antes de elegir su carrera, además, que durante los primeros meses de ingreso a la universidad se tomen acciones encaminadas a prevenir la deserción (8).

Por otra parte, se puede describir al abandono estudiantil como un fenómeno complejo, el cual involucra variedad de características de índole social (particularidades familiares, comunitarias, singularidades personales), político (leyes, reglamentos, decretos), económico (ingresos familiares, ingresos propios, clase social) y académico (modelos pedagógicos, becas, metodología aplicada por los profesores, currículo, todos estos factores constituyen un riesgo para una posible deserción, en consecuencia, es menester la participación activa de docentes, estudiantes, y demás acciones que contribuyan a disminuir los factores de riesgo (9).

La deserción se da por lo general en los niveles iniciales de la carrera (10), y lo provocan factores de tipo psicológico (rasgos de personalidad, creencias, actitudes, sociológico (escasa integración con el entorno), organizacional (características del docente, particularidades de la universidad, experiencias en el entorno educativo), esta problemática trae consecuencias sociales y económicas, acarreando desempleo, pobreza, que impide mejorar la calidad de vida de los estudiantes, la mayoría de los estudiantes que abandonan sus estudios, se da por motivos académicos, se considera al promedio como el factor más significativo para la deserción, sin embargo la edad no es menos significativa la mayoría de los estudiantes que deciden abandonar sus estudios, tienen menos de 20 años y en un gran porcentaje son hombres que pertenecen al programa de matemáticas (11).

La regresión binaria, se emplea principalmente en la medicina para medir la relación existente entre una variable objetivo y una o más variables explicativas (12), mediante la aplicación de estadísticos, que conllevan al análisis y descripción del fenómeno estudiado (13).

La predicción basada en un modelo logit (14), para identificar determinantes del abandono escolar, toma como variable dependiente a la deserción e independientes al género, edad, nivel de posgrado de la familia, rendimiento académico del estudiante, estado civil, número de hijos, tipo de beca o financiamiento para cubrir el costo de los estudios. Con los datos de 1294 estudiantes, se empleó la regresión logística para desarrollar un modelo de abandono escolar, clasificando la variable dependiente como “1” igual a “el alumno interrumpió sus estudios”, y “0” como “el alumno no interrumpió sus estudios”. Las categorías de las variables independientes se definieron como variables dicotómicas. El modelo muestra una tasa de predicción del 97,1%, indicando que, de los encuestados, el 97.1% se clasificó correctamente en categorías de la variable dependiente.

Es posible predecir el rendimiento académico, tomando en cuenta variables significativas como la asistencia y la participación activa de los estudiantes en el aula clase. Con una muestra de 175 estudiantes de la asignatura de “Métodos y Diseños de Investigación en Psicología”, se crea un modelo agregando datos sobre la nota del examen final, asistencia y participación en clases, puntuación del bachillerato, cuestionarios, nota con la cual ingreso a la universidad, especialidad de cognitiva-metodología, motivación al iniciar la carrera y estimación por la vocación(15).

Se emplea la regresión lineal múltiple para contrastar el efecto que produce la asistencia y la participación activa en el pronóstico del rendimiento académico y la regresión logística para identificar las variables que mejor se acoplen a la predicción del éxito/fracaso académico. Para pronosticar el éxito o fracaso académico, se realiza

un análisis previo de los datos, logrando categorizar la variable dependiente u objetivo en 90 alumnos aprobados, y 85 suspensos.

Un modelo de regresión logística analiza la relación entre las variables independientes y dependiente que influyen en el rendimiento académico, así como su grado de significancia e incidencia en el modelo. Se toma una muestra de 287 estudiantes, durante el periodo académico 2003 - 2007. Las variables utilizadas son: promedio en la universidad (PromU), promedio del bachillerato (PromBR), carrera, sector donde se encentra ubicado el plantel (SectorP), eficiencia (Efi), edad, turno del plantel (diurno, nocturno) y sexo (16). Se construye un modelo de posibilidades proporcionales o logit acumulado, utilizando la regresión logística múltiple para las variables objetivos PromU, Efi y las variables independientes SectorP y PromBR con dos niveles o categorías, logrando conseguir un alto nivel de probabilidad en la predicción.

Por otra parte (17), desarrollan un modelo de regresión logística para la predicción del rendimiento estudiantil, consideran la variable dependiente: rendimiento con tres criterios (c1: aprobar 3 o más materias, c2: aprobar 4 o más materias, c3: aprobar todas las materias), y a las variables independientes: Expectativa (Ex), valencia (VA), notas de enseñanza media (PtNt), instrumentalidad (INS), puntaje en matemáticas (PtMt), puntaje en lenguaje (PtLg), puntaje en ciencias (PtCs), se obtienen los coeficientes de la ecuación a través del estadístico de WALD, utilizando el software Statgraphics Plus, se verifica que las variables PtNt, PtMt tienen un p_valor $< 0,01$, por lo cual se considera, como variables significativas del modelo. Tomando en consideración el primer semestre de la carrera se deduce, en base al c1, que el modelo de regresión logística tuvo una probabilidad del 71,4% de éxito en la predicción, al igual que el c2, e incrementa a 75% para el c3. De los estudiantes que el modelo predijo como no exitosos (fracaso), usando el c1 el 78,23% no tuvo éxito, el valor se incrementa a 85,2% con el c2 y aumenta a 92,34% con el c3, se concluye, que el modelo de regresión logística predice exitosamente a la mayoría de los sujetos si se toma en consideración el c1, puesto que el estudio llevó a cabo también un análisis comparando los dos modelos resulta una diferencia equilibrada en la predicción.

El modelo de regresión logística que se usa para pronostica el promedio de calificaciones (GPA), en el primer semestre de la Facultad de Ingeniería de la Información, usa una tabulación cruzada entre los datos categóricos de predicción y el valor de los datos categóricos de observación. Se trabaja con datos de estudiantes desde el 2008 al 2015. Las variables utilizadas son: rendimiento académico de la escuela secundaria, habilidad de pruebas numéricas, verbales, espaciales y de analogías (18).

Se desarrolla un modelo predictivo para verificar cada una de las variables que ayudó a predecir el interés de los estudiantes de secundaria en especializarse en

ciencias, matemáticas, ingeniería y tecnología. Las variables tanto dependientes como independientes utilizadas en el estudio son: género, raza, rendimiento académico en asignaturas como matemáticas, inglés, ciencias, se tomó datos de tipo familiar como: ingresos familiares, número de hermanos, expectativas de obtener un trabajo durante el desarrollo de sus estudios, tipo de programa de segundaria del cual proviene, asistencia, movilidad, promedio de notas, todas la variables fueron clasificadas como nominal, escalar, ordinal, dicotómicas (19). Para la construcción del modelo se levantan datos del año 2003 de 59618 estudiantes 53% mujeres y 47% hombres, la mayoría de los estudiantes eran blancos (74%) y de un ingreso familiar bajo o medio bajo (52%). Las estadísticas descriptivas y predictivas se calcularon en función de una muestra aleatoria utilizando SPSS, se creó un modelo de regresión logística binomial de un solo nivel que permita predecir el interés temprano por especializarse en ciencias, matemáticas, ingeniería y tecnología.

Los elevados índices de deserción estudiantil en las universidades del Ecuador, afecta la gratuidad de la educación. La Carta Magna del Ecuador en su sección quinta Art. 28 manifiesta: “La educación pública será universal y laica en todos sus niveles, y gratuita hasta el tercer nivel de educación superior…” (20).

En el Ecuador en promedio se matriculan 700.000, estudiantes en el nivel superior, según datos que muestra la Secretaría Nacional de Educación Superior Ciencia y Tecnología (SENECYT), el 76% del total logra concluir sus estudios universitarios, mientras que el 26% , son estudiantes que desertan durante el periodo que dura la carrera (21). Las deserciones se dan principalmente durante los primeros niveles. En países como Costa Rica, Ecuador, los porcentajes de deserción van entre el 25% y 35%, lo que se constituye en un problema para los países en vías de desarrollo (22). La UTA, al igual que las diferentes universidades del país atraviesa por el mismo fenómeno de deserción estudiantil, escasos datos existen sobre la totalidad de desertores de las diferentes facultades, es poco evidenciable las causas que genera el fenómeno.

Siendo la deserción un fenómeno que afecta aproximadamente al 25%, de la población universitaria del país (21), y sabiendo que acarrea gastos significativos para el estado, es menester desarrollar medidas que mitiguen los efectos que produce sobre los estudiantes y la sociedad en general, a través del uso de técnicas predictivas.

El propósito principal es predecir a futuros desertores con un nivel de probabilidad aceptable, generando un modelo predictivo de deserción estudiantil usando regresión logística binaria, de esta manera la Institución de Educación Superior podrá tomar medidas preventivas para evitar una posible deserción. Para alcanzar este propósito, se debe analizar las variables que provocan la deserción estudiantil, así como los factores que inciden en la decisión de los estudiantes de mantenerse en la carrera

seleccionada, teniendo en cuenta aspectos de tipo social, económico, sociodemográfico y de rendimiento académico de los estudiantes, al identificar y analizar dichos factores se aplica la regresión logística binaria para predecir a los posibles desertores.

El desarrollo de un modelo predictivo de deserción estudiantil, requiere llevar a cabo acciones como la recolección y análisis de datos, a través de la exploración exhaustiva de archivos históricos que almacenan registros de los estudiantes sobre el periodo académico de ingreso a la universidad, cursos en los cuales se encuentran matriculados, materias que reciben, promedio de notas de diferentes asignaturas, notas obtenidas en el examen de suspensión, además de datos de tipo etnográfico como ciudad de nacimiento y de residencia (Ambato/otros), género (masculino/femenino), estado civil (soltero/ casado/ divorciado/ unión libre), etnia o raza y fecha de nacimiento. Por otra parte, es necesario el uso de una técnica multivariante como la regresión logística binaria múltiple para identificar las variables relevantes que inciden en la deserción, así como los coeficientes de la ecuación matemática.

Bajo este contexto se proponen varias estrategias encaminadas a disminuir los índices de deserción, concientizar a las autoridades y fomentar procesos de detección temprana de la problemática, se plantean alternativas en la creación de modelos predictivos, unos basados en la aplicación de técnicas de minería de datos y otros en el desarrollo de modelos matemáticos, poniendo énfasis en la regresión logística para predecir las categorías de los estudiantes (18).

2. Materiales y Métodos

La metodología de trabajo, se divide en cuatro fases partiendo del análisis de la información histórica de una base de datos de 425 estudiantes con 15474 registros, pertenecientes a la Carrera de Docencia en Informática de la UTA, que permite detectar y analizar las variables que intervienen en el modelo, posteriormente se procede a la formulación del modelo a través de ecuaciones matemáticas, seguido de la evaluación del modelo (Figura 1).

Figura 1

Etapas de la metodología.

2.1. Modelo de regresión logística binaria múltiple

Este modelo se utiliza cuando se tiene dos o más variables explicativas que pueden ser cualitativas o cuantitativas y que a diferencia de la regresión lineal múltiple no es indispensable que se verifique el supuesto de normalidad (23).

Por lo general este modelo emplea el método de máxima verosimilitud, para estimar los coeficientes del modelo a través de iteraciones que van ajustando el modelo (24).

La formulación matemática de la curva logística es:

$$f(z)=\frac{1}{1+e^{z}}, \quad \infty- < z < \infty+, \tag{1}$$

Donde $f(z)$ puede expresarse como una aproximación lineal de:

$$x_1, x_2, x_3 \ldots, x_k,$$

de modo que:

$$z=\alpha+\beta_1 x_1+ \beta_2 x_2+ \cdots + \beta_k x_k = (\alpha+\sum \beta_i x_i), \tag{2}$$

de forma equivalente:

$$y=f(z)=\frac{1}{1+e^{-(\alpha+\sum \beta_i x_i)}} \tag{3}$$

Los valores de α y β_i son parametros desconocidos y se estimarán a partir del conjunto de datos en forma de muestras utilizando el estimador de máxima verosimilitud.

Este modelo está diseñado para describir la probabilidad que siempre se encuentra entre 0 y 1 (25). Por tanto la regresión logística, emplea la técnica de máxima verosimilitud, a través de un proceso repetitivo, compuesto de fases sucesivas que regulan el modelo (24).

El resultado de la función logística, es un valor de probabilidad que pronostica una acción y facilita la interpretación de resultados, que son generados por el modelo de regresión logística, que resulta como parte de la asignación de valores a las variables explicativas seleccionadas para forma parte de modelo y que generan valores de salida en la variable dependiente. El modelo de regresión logística binaria múltiple contempla, básicamente dos eventos, representados como 0 y 1 (éxito y fracaso). En términos de probabilidad se puede expresar como: la probabilidad de ocurrencia de un evento representado por P, y la probabilidad de no ocurrencia de un evento expresado por $1-P$:

probabilidad que un evento sea exitoso:

$$Pr\,(y=1) = P\ ,$$

probabilidad que un evento no suceda

$$Pr\,(y=0) = 1-P\ .$$

Puesto que se trata de una regresión logística binaria múltiple es necesario considerar la información de más de una variable explicativa para el desarrollo de un modelo que pueda predecir el valor de la variable objetivo y.

2.2. Análisis variables

En esta fase se realizó la recolección e integración de datos, que pertenece a 425 estudiantes. La institución proporciono los datos crudos a partir del año 2006 en dos matrices que fueron integradas, para su posterior procesamiento y limpieza, a través de la cual se detectó datos atípicos en diferentes variables, mismos que fueron eliminados obteniendo datos limpios. La información obtenida de los estudiantes contiene datos sobre: edad, nota promedio de las asignaturas tomadas por el estudiante en el primer parcial del primer semestre, nota promedio de las asignaturas tomadas por el estudiante en el segundo parcial del primer semestre, nota promedio de las asignaturas del primer parcial del segundo semestre, nota promedio de las asignaturas tomadas por el estudiante en el segundo parcial del segundo semestre. Además, datos de tipo etnográficos como: género, estado civil, etnia, fecha de nacimiento, ciudad de nacimiento y de residencia, la totalidad de la información consta de 15474 registros.

Las variables con las que se cuenta para la generación del modelo se clasificaron (Tabla I) (26) en:

1. Nominal (los valores son básicamente una función de etiquetado).
2. Ordinal (los valores obedecen a la relación de orden).
3. Cuantitativo (los valores tienen todo el poder expresivo de los números reales).

Se tomó únicamente las notas promedio del primero y segundo semestre por ser los niveles en los cuales existen mayor cantidad de estudiantes desertores.

En concreto se trabaja con 10 variables independientes, detalladas en la (Tabla I) y una dependiente expuesta en la (Tabla II).

Tabla 1

Clasificación de las variables independientes para la generación del modelo.

Variables independientes	Tipo	Categorías
Género (G)	Cualitativa – Nominal dicotómica	Masculino Femenino
Estado Civil (EC)	Cualitativa – Nominal politómica	Soltero Casado Divorciado Unión libre
Raza (RA)	Cualitativa – Nominal politómica	Mestiza Indígena Blanca
Edad (ED)	Cuantitativo	
Lugar de Nacimiento (LN)	Cualitativa – Nominal dicotómica	Ambato Otros
Ciudad de residencia (CR)	Cualitativa – Nominal dicotómica	Ambato Otros
Promedio del primer parcial del primer semestre (Nota1p: 1= primer parcial, p = primer semestre)	Cuantitativo	
Promedio del segundo parcial del primer semestre (Nota2p: 2= segundo parcial, p = primer semestre)	Cuantitativo	
Promedio del primer parcial del segundo semestre (Nota1s: 1= primer parcial, s = segundo semestre)	Cuantitativo	
Promedio del segundo parcial del segundo semestre (Nota2s: 2= segundo parcial, s = segundo semestre)	Cuantitativo	

Tabla 2

Variable dependiente.

Variable Dependiente	Tipo	Categorías
Desertor	Cualitativa - Nominal dicotómica	No Desertor = 0 Desertor = 1

2.3. Formulación del modelo

Para el desarrollo del modelo predictivo de deserción estudiantil basado en regresión logística binaria múltiple, se toma como base, los datos de 425 estudiantes y se procede a obtener los coeficientes de las variables independientes o explicativas.

Entonces, la creación del modelo de regresión logística se lleva a cabo usando el programa SPSS, partiendo del siguiente caso práctico: Se desea predecir si un estudiante es desertor (categorizando la variable dependiente en: No desertor, desertor), partiendo de variables independientes: G, EC, RA, ED, LN, CR, Nota1p, Nota2p, Nota1s, Nota2s, que permitan predecir la deserción de un estudiante.

Se procede a la aplicación del método de máxima verosimilitud, para obtener una combinación lineal de las variables explicativas definidas para el modelo, G, EC, RA, ED, LN, CR, Nota1p, Nota2p, Nota1s, Nota2s, el procedimiento aplicado permitirá calcular la probabilidad, que clasificará en las categorías establecidas para la variable objetivo (Desertor o No desertor). Por ello, se plantea la pregunta:

¿Están relacionados el género, el estado civil, la raza, la edad, el lugar de nacimiento, la ciudad de residencia, nota1p, nota2p, nota1s, nota2s, con el hecho de que un estudiante sea desertor?

Para ejecutar el análisis de regresión logística binaria se utilizó el software SPSS, se aplica el estadístico de Wald hacia delante, y en función de las variables explicativas se determina los coeficientes de la ecuación, además de definir cuál de estas variables resultan significativas para el modelo y cuales se descartan. El método de Wald, es un proceso automático, conformado de un conjunto de pasos hacia adelante, que se emplea para verificar cuál de las covariables, deben incluirse o excluirse en el modelo. Una de las ventajas de usar este método, es que el investigador, no interviene en la inclusión o exclusión de las variables del modelo, pues es el estadístico quien se encarga de ir introduciendo y a la vez descartando las variables en cada iteración. Se parte de un modelo que no contiene ninguna covariable, en el primer paso se introduce una variable y se analiza su grado de significancia, este proceso se repite hasta incluir a todas las variables de estudio en cada paso se va descartando las variables poco significativas.

Es necesario mencionar que los “métodos automáticos por pasos” son apropiados para obtener una variedad de modelos, con un propósito predictivo (23). Previo a realizar la interpretación del modelo es necesario considerar determinadas salidas del análisis de regresión logística:

El análisis tomo en cuenta a 425 casos, no existe ningún caso perdido o con datos anómalos como lo muestra la (Tabla III).

Tabla 3

Variable dependiente.

Casos no ponderados		N	Porcentaje
Casos seleccionados	Incluidos en el análisis	425	100,0
	Casos perdidos	0	,0
	Total	425	100,0
Casos no seleccionados		0	,0
Total		425	100,0

Se asigna 0 a la categoría NO desertor y 1 a la categoría Desertor, según se expone en la (Tabla IV).

Tabla 4

Codificación de la variable dependiente.

Valor original	Valor interno
No Desertor	0
Desertor	1

La (Tabla V), indica la codificación de las variables categóricas así: la variable estado civil tiene tres categorías, la categoría unión libre es el valor de referencia, mientras que se codifica casado (1), divorciado (2), soltero (3), de la variable género se toma como referencia a la categoría masculino y se codifica a la categoría femenino (1), no se toma en cuenta el resto de variables por no ser significativas para la formulación del modelo.

Tabla 5

Codificación de las variables categóricas.

		Frecuencia	Codificación de parámetros		
			(1)	(2)	(3)
Estado Civil	Casado (1)	65	1,000	,000	,000
	Divorciado (2)	1	,000	1,000	,000
	Soltero (3)	357	,000	,000	1,000
	Unión libre (Valor de referencia)	2	,000	,000	,000
Raza	Blanca (1)	1	1,000	,000	
	Indígena (2)	17	,000	1,000	
	Mestiza (Valor de referencia)	407	,000	,000	
Lugar Nacimiento	Ambato (1)	263	1,000		
	Otros (Valor de referencia)	162	,000		
Ciudad Residencia	Ambato (1)	316	1,000		
	Otros (Valor de referencia)	109	,000		
Genero	Femenino (1)	183	1,000		
	Masculino (Valor de referencia)	242	,000		

Se presenta los cálculos e interpretación del modelo:

Tabla 6

Parámetros estimados por el modelo – Variables en la ecuación.

		B	E.T.	Wald	gl	Sig.	Exp (B)	I.C. 95% para EXP (B)	
								Inferior	Superior
Paso 5	Género (1)	-,874	,336	6,753	1	,009	,417	,216	,807
	Estado Civil			30,809	3	,000			
	Estado Civil (1)	3,572	8,112	,194	1	,660	35,603	,000	2,858E8
	Estado Civil (2)	-15,907	40192,971	,000	1	1,000	,000	,000	.
	Estado Civil (3)	1,411	8,105	,030	1	,862	4,100	,000	32492489,838
	Edad	-,322	,084	14,556	1	,000	,725	,615	,855
	Nota1s	-,487	,237	4,228	1	,040	,614	,386	,977
	Nota2s	-,679	,157	18,661	1	,000	,507	,373	,690
	Constante	12,255	8,400	2,128	1	,145	209949,507		

2.3.1. Bloque 1: Método = Por pasos hacia adelante (Wald)

En relación a los resultados obtenidos se interpreta, que en la ecuación de regresión logística binaria solamente aparecerán las variables nota2s, estado civil, edad, género y nota1s (Tabla VI). Quedan fuera las variables: nota1p, nota2p, lugar de nacimiento, ciudad de residencia y raza (Tabla VII), para el análisis se toma en cuenta el paso 5.

Tabla 7

Variables que no están en la ecuación.

			Puntuación	gl	Sig.
Paso 5	Variables	Raza	1,423	2	,491
		Raza (1)	,114	1	,736
		Raza (2)	1,316	1	,251
		Lugar Nacimiento (1)	,013	1	,908
		Ciudad Residencia (1)	,286	1	,593
		Nota1p	1,932	1	,165
		Nota2p	,216	1	,642
	Estadísticos globales		3,566	6	,735

2.3.2. Método de Wald

Cuando se requiere estudiar únicamente la significación de un determinado parámetro B_k, es decir sólo, se requiere conocer si se cumple la hipótesis $B_k = 0$, se puede utilizar el método de Wald. El test de Wald consiste en dividir el valor estimado b_k, por su

desviación típica estimada s(b_k), La hipótesis $B_k = 0$, se rechaza si el valor obtenido es mayor que el percentil $1-$ α, donde α es el riesgo que esté dispuesto asumir.

La (Tabla VII), muestra el procedimiento que el método Wald ha realizado para seleccionar las variables del modelo, en el paso 1, ha escogido la variable nota2s, a través del estadístico "Puntuación eficiente de Rao". En el paso 2, de entre las candidatas ha seleccionado la variable estado civil. Posteriormente, en el paso 3, ha introducido a la variable edad, en el paso 4, ha elegido la variable género. Finalmente, en el paso 5 de la (Tabla VII), de las variables sobrantes ha seleccionado a la variable nota1s, en este paso a través del estadístico Wald las variables eliminadas de las 10 seleccionadas han sido nota1p, nota2p, lugar de nacimiento, ciudad de residencia y raza. Con los datos de la (Tabla VII), se construye la ecuación de regresión logística binaria.

Reemplazando los coeficientes β en (2):

$$y = 12,255 - 0,679Nota2s + 3,572\ EC - 0,322ED - 0,874G - 0,487Nota1s\ . \qquad (4)$$

A partir de la ecuación logística, podemos determinar la probabilidad que un estudiante sea desertor o no deserto, es decir si un estudiante va a desertar, teniendo en cuenta la nota promedio del segundo parcial del segundo semestre (nota2s), su estado civil, edad, género, y la nota promedio del primer parcial del segundo semestre (nota1s).

$$y = 12,255 - 0,679\,(1) + 3,572\,(1) - 0,322\,(1) - 0,874\,(1) - 0,487\,(1) \ = \ 13.465$$

Reemplazando en la expresión (1):

$$P(d) = \frac{1}{1+e^{-13.465}} = 0,999 \qquad (5)$$

donde d es desertor.

Se ha clasificado con 0 a no desertor y 1 a desertor. En consecuencia, la probabilidad que un alumno cuente con una nota2s, su estado sea igual a casado, tenga una determinada edad, sea de género femenino, y haya obtenido una nota1s, el modelo lo clasificara como desertor, con una probabilidad de certeza del 99,9%.

La (Tabla VIII), muestra la matriz de correlación del modelo de regresión, a partir de ésta se realiza un sondeo sobre la presencia de multicolinealidad entre las variables explicativas que forman parte del modelo. Si los valores de la correlación son altos entonces se deduce que existe multicolinealidad entre las variables caso contrario que no.

Por tanto, con una correlación fuerte, la incidencia de la variable explicativa, sobre la variable objetivo, podría estar sesgada, es decir, las dos variables pueden contener la misma información por lo cual no aportarían a la predicción, pues sostienen una relación con otras variables.

Para el caso de estudio las correlaciones son relativamente bajas, sin embargo, no se puede destacar la correlación alta que muestra la constante y la variable estado civil, lo que podría implicar un problema de multicolinealidad.

Tabla 8

Matriz de Correlaciones.

		Constant	Género (1)	Estado Civil (1)	Estado Civil (2)	Estado Civil (3)	Edad	Nota1s	Nota2s
Paso 5	Constant	1,000	-,065	-,949	,000	-,953	-,262	-,117	-,018
	Género (1)	-,065	1,000	,003	,000	,009	,186	,023	-,003
	Estado Civil (1)	-,949	,003	1,000	,000	,999	-,007	-,061	-,005
	Estado Civil (2)	,000	,000	,000	1,000	,000	,000	,000	,000
	Estado Civil (3)	-,953	,009	,999	,000	1,000	,001	-,061	,002
	Edad	-,262	,186	-,007	,000	,001	1,000	,236	,034
	Nota1s	-,117	,023	-,061	,000	-,061	,236	1,000	-,591
	Nota 2s	-,018	-,003	-,005	,000	,002	,034	-,591	1,000

2.4. Evaluar el modelo

El modelo creado, generó una función de regresión logística que predice con un nivel de probabilidad aceptable, la clasificación de la variable dependiente en función de las variables dependientes. Utilizando los coeficientes Pseudo R2, las tablas de clasificación o matriz de confusión, y el contraste de bondad de ajuste de Hosmer y Lemeshow, se evalúa el modelo en función de los valores obtenidos con estos estadísticos.

2.4.1. Coeficientes de determinación Pseudo R2

La (Tabla IX), muestra el resumen del modelo, con cinco iteraciones de solución del modelo. Se observa que los valores de verosimilitud, R2 de Cox y Snell, y R2 de Nagelkerke, mejoran a medida que el modelo se va depurando. El valor del "log de la verosimilitud", se va reduciendo, indicando, que se va maximizando con cada iteración.

Tabla 9

Resumen del modelo.

Paso	-2 log de la verosimilitud	R^2 de Cox y Snell	R^2 de Nagelkerke
5	262,901	,472	,660

El estadístico R2 de Cox y Snell no puede alcanzar el valor 1, sin embargo, es necesario que su valor se acerque a 1 para mejorar la predicción. Por otra parte, si conviene que el valor del estadístico R2 de Nagelkerke llegue alcanzar el valor 1, pues señala la existencia de una regresión perfecta (27).

Las ecuaciones de estos estadísticos son:

R2 de Cox y Snell,

$$R^2 = 1 - \left[\frac{-2LL_{null}}{-2LL_k}\right]^{\frac{2}{n}}.$$

R2 de Nagelkerke,

$$R^2 = \frac{1 - \left[\frac{-2LL_{null}}{-2LL_k}\right]^{2/n}}{1 - (-2LL_{null})^{2/n}},$$

LL_{null} , es el Log likelihood del modelo (log de la verosimilitud),

LL_k, es el log likelihood que integran las variables restantes en el modelo.

En el caso de estudio el valor de R2 de Cox y Snell = 0,472 ≡ 47,2%, es un valor menor que 1, lo que valida la precisión de la predicción. Se puede decir, que el modelo predice correctamente las categorías con un nivel de certeza del 47,2%.

R2 de Nagelkerke = 0,660 ≡ 66%, un valor relativamente aceptable, pues debe ser lo más cercano a 1. Se deduce entonces, que el modelo puede predecir con un nivel de certeza del 66%, resultando más significativo el uso de este estadístico.

Por otra parte, si tómanos en cuenta los valores de R2 de Cox y Snell y R2 de Nagelkerke (Tabla IX), podemos determinar en qué medida se puede predecir los resultados de la deserción, considerando las variables seleccionadas para la creación del modelo (Tabla VI).

2.4.2. Tablas de clasificación – Matriz de confusión

La (Tabla X), muestra que, de un total de 425 estudiantes, 371 han sido clasificados correctamente y 54 se ha clasificado incorrectamente. La probabilidad de certeza que el modelo tiene para clasificar tanto a desertores como a no desertores es del 87,3% con un valor de: $\propto = 0,5$, como valor de corte.

Tabla 10

Tabla de clasificación.

Observado			Pronosticado		
			Desertor		Porcentaje correcto
			No Desertor	Desertor	
Paso 5	Desertor	No Desertor	276	12	95,8
		Desertor	42	95	69,3
	Porcentaje global				87,3

2.4.3. Análisis de la prueba Ómnibus

Para analizar este estadístico se debe verificar que el valor de significancia sea menor a 0.05, entonces se podrá deducir que el modelo es adecuado, es decir, las variables independientes explican a la variable dependiente.

La (Tabla XI), muestra la prueba omnibus que se ha depurado para cada uno de los cinco pasos del modelo. Los resultados indican que hasta la iteración 3, los coeficientes están bien seleccionados, y las variables que forman parte del modelo tienen valores diferentes a 0, sin embargo, en la iteración 4 y 5 los valores tienden a cambiar. En cuanto a verificar si las variables seleccionadas para el desarrollo del modelo, predicen o no predicen la deserción, se confirma a través de la prueba omnibus.

El análisis se centra en el p_valor bajo la siguiente condición "si p_valor de la prueba omnibus es menor que 0,05 ($p < 0,05$), diríamos que las variables seleccionadas SI pueden predecir la deserción mediante el modelo de la regresión logística binaria". Para el caso de estudio $Sig. = 0,000$ y el $p < 0,05$, por tanto, las variables independientes seleccionadas para crear el modelo SI predicen la deserción.

Tabla 11

Pruebas omnibus sobre los coeficientes del modelo.

		Chi cuadrado	gl	Sig.
Paso 5	Paso	4,289	1	,038
	Bloque	271,435	7	,000
	Modelo	271,435	7	,000

2.4.4. Contraste de bondad de ajuste de Hosmer y Lemeshow

La (Tabla XII), muestra la prueba de bondad de ajuste del modelo con un nivel de significación de 5%, se puede decir que el modelo es relativamente adecuado, pues se ajusta a los datos, $p\_valor = 0,534$.

Tabla 12

Prueba de Hosmer y Lemeshow.

Paso	Chi cuadrado	gl	Sig.
5	10,468	8	,534

La (Tabla XIII) muestra la tabla de contingencia que describe la asociación entre las variables del modelo, para cada iteración:

Tabla 13

Tabla de contingencia para la prueba de Hosmer y Lemeshow.

		Desertor = No Desertor		Desertor = Desertor		Total
		Observado	Esperado	Observado	Esperado	
Paso 5	1	42	42,309	1	,691	43
	2	37	40,088	5	1,912	42
	3	40	40,139	3	2,861	43
	4	39	37,319	2	3,681	41
	5	39	37,632	4	5,368	43
	6	40	35,689	3	7,311	43
	7	30	31,731	13	11,269	43
	8	19	20,242	24	22,758	43
	9	2	2,818	41	40,182	43
	10	0	,032	41	40,968	41

Se responde entonces a la pregunta propuesta:

¿Están relacionados el género, el estado civil, la raza, la edad, el lugar de nacimiento, la ciudad de residencia, nota1p, nota2p, nota1s, nota2s, con el hecho de que un estudiante sea desertor?

La respuesta se da en función del análisis del modelo de regresión logística, las variables que están relacionadas con la predicción de la deserción son nota2s, estado civil, edad, género y nota1s, y las que no tienen relación son: Nota1p, nota2p, lugar de nacimiento, ciudad de residencia y raza.

3. Desarrollo y discusión

Se realiza una breve descripción de modelo obtenido, se verifica el cumplimiento de supuestos y los resultados obtenidos posterior a la aplicación del modelo.

3.1. Descripción del modelo

Para llevar a cabo el proceso de análisis predictivo, se requiere seguir diferentes etapas como: recolectar, ordenar y adaptar los datos, analizarlos, elegir un modelo matemático, optimizar los parámetros. En función de lo expuesto se procede a describir cada una de las etapas que se cumplieron para el desarrollo del modelo de regresión logística binaria.

Se seleccionan 425 datos de entrenamiento para estimar los coeficientes del modelo predictivo, las variables dependientes (género, estado civil, raza, edad, lugar de nacimiento, ciudad de residencia, nota1p, nota2p, nota1s, nota2s), además se identifica la variable dependiente como desertor y se codifica:

$$(No\ desertor = 0,\ \ Desertor = 1).$$

Se aplica el estadístico Wald por pasos hacia delante disponible en el software SPSS, para estimar los parámetros del modelo, con cinco iteraciones las variables seleccionada son nota2s, estado civil, edad, género, y nota1s, las variables que no constan en la ecuación son raza, lugar de nacimiento, ciudad de residencia, nota1p, nota2p.Con las variables seleccionadas se construye la ecuación de regresión logística:

$$y = 12,255 - 0,679Nota2s + 3,572\ EC - 0,322ED - 0,874G - 0,487Nota1s.$$

Por medio de la matriz de correlación se detecta que existe una correlación alta entre la constante y la variable estado civil, lo que puede implicar un problema de multicolinealidad. La evaluación del modelo se hace a través de la tabla de clasificación que muestra de un total de 425 estudiantes, 371 han sido clasificados de manera correcta, teniendo al 87,3% de la población correctamente clasificada como No desertor o desertor.

A través de los estadísticos R2 de Cox y Snell y R2 de Nagelkerke, se determina la medida para la predicción de los resultados de la deserción, considerando las variables elegidas para la creación del modelo. En cuanto a verificar si las variables escogidas para el desarrollo del modelo, predicen o no predicen la deserción, se confirma a

través de la prueba omnibus que, para el caso de estudio, $p\_valor < 0,05$, por tanto las variables explicativas seleccionadas para crear el modelo SI predicen la deserción.

Además, por medio del estadístico R2 de Cox y Snell se constata que el modelo tiene un nivel de certeza en sus predicciones del 47,2%, lo cual no resulta significativo al momento de la predicción, sin embargo, el estadístico R2 de Nagelkerke muestra una capacidad de predicción del 66%, resultando más apropiado el uso de este estadístico.

3.2. Cumplimiento de supuestos

Se determinan los supuestos, con el propósito de conocer si la solución encontrada a través de la aplicación de la regresión logística binaria múltiple en la generación del modelo predictivo es estable.

3.2.1. Linealidad

Este supuesto se centra en la existencia de una relación lineal entre las variables predictoras y la variable dependiente.

1. Primero, se calcula la correlación bivariada de las variables cuantitativas y cualitativas. Las variables categóricas han sido previamente transformadas en variables indicadoras (dummy), para aplicar la correlación. La (Tabla XIV) muestra la correlación existente entre las variables predictoras y la variable resultado del modelo.

Tabla 14

Correlación bivariada.

Correlación de Pearson		Desertor
Genero	Correlación de Pearson	,173**
	Sig. (bilateral)	,000
Estado Civil	Correlación de Pearson	-,222**
	Sig. (bilateral)	,000
Edad	Correlación de Pearson	-,124*
	Sig. (bilateral)	,010
Nota1s	Correlación de Pearson	-,609**
	Sig. (bilateral)	,000
Nota2s	Correlación de Pearson	-,676**
	Sig. (bilateral)	,000

**. La correlación es significativa al nivel 0,01 (bilateral).
*. La correlación es significante al nivel 0,05 (bilateral).

Los resultados muestran que existe una correlación directa de 0,173, entre la variable género y la variable desertor a pesar de ser baja es significativa por tanto, se mantendrá en el modelo, por otra parte la relación entre desertor y estado civil es de -0,222, a pesar de ser negativa sigue siendo significativa para el modelo, de igual entre la variable desertor y edad existe una correlación inversa de -0,124. Finalmente, la correlación más alta se obtiene entre las variables nota1s y desertor de -0,609 y entre nota2s y desertor de -0,676 lo que significa que son altamente significativas para el modelo.

En tal virtud, las variables: género, estado civil, edad, nota1s, nota2s, tienen una correlación lineal significativa con la variable desertor, por tanto, se mantienen en el modelo de deserción estudiantil.

1. Segundo, se toma en consideración lo expuesto por (28), sobre el uso del registro (logit) con el fin de dar cumplimiento al supuesto de linealidad, para lo cual es preciso la existencia de una relación lineal entre alguna variable independiente continua y el logit de la variable dependiente. Para evidenciar este supuesto se requiere conocer si el valor de la interacción entre la variable predictora y su transformación logarítmica es significativo.

Es necesario ejecutar la regresión logística incorporando como variables predictoras las iteraciones entre cada predictor y el logaritmo de sí mismo (ver código fuente Anexo 2), se obtienen los logaritmos de las variables edad, nota1s y nota2s, no se toma en cuenta para la obtención del logaritmo a las variables género y estado civil por ser variables indicadoras (dummy), sin embargo, si se las incluye al ejecutar la regresión logística. Los valores de los logaritmos de las variables edad, nota1s y nota2s se almacenan en logedadlnt, lognota1slnt, lognota2slnt, respectivamente (Tabla XV). Para realizar el contraste se integran todas las variables iniciales y se incluyen las nuevas iteraciones.

Tabla 15

Iteraciones.

Coeficientes	Std. Error	z value	Pr(>|z|)
Edad	0.12781	0.696	0.48672
Nota1s	7.27520	1.703	0.08865.
Nota2s	5.38590	-2.388	0.01695 *
logedadlnt	0.03729	-2.754	0.00588 **
lognota1slnt	2.48321	-1.795	0.07265.
lognota2slnt	1.85011	2.315	0.02059 *

Significado de los códigos: 0 '***'0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Número de iteraciones de puntuación de Fisher: 8

De la (Tabla XV), se analiza el valor de Pr (>|z|). Entonces si Pr (>|z|) >0.05 los valores de las iteraciones son significativos (28). Por tanto, las variables pueden formar parte del modelo. Se indica entonces que el supuesto de linealidad se cumple para la variable edad, nota1s, se debe tener en cuenta la variable nota2s pues su valor es relativamente bajo, lo que no excluye su significancia para el modelo, como los demuestra la correlación expuesta en la (Tabla XIV). Se analizado el supuesto de linealidad desde dos enfoques, por medio de la correlación bivariada y a través del registro logit confirmando que se cumple el supuesto de linealidad, por tanto, las variables que conforman el modelo predictivo de deserción estudiantil aportan significativamente.

3.2.2. Multicolinealidad

Para comprobar el supuesto partimos de la premisa que se reduce el poder predictivo de las variables independientes si dos o más variables tienen una correlación fuerte, pues significa que hay más de dos variables predictoras aportando la misma información. La (Tabla VIII), muestra la matriz de correlación en la cual se evidencia que:

Género y estado civil tienen correlación débil con las variables nota2s, edad, género y nota1s. La variable edad, estado civil, nota2s, género y nota1s, muestran una correlación débil. De igual manera la variable nota1s, estado civil, edad y género tienen correlación débil, sin embargo, se puede evidenciar que la correlación se eleva entre las variables nota1s y nota2s. La variable nota2s tiene escasa relación con estado civil, edad y género. Al igual que en el análisis anterior la correlación entre la variable nota2s y nota1s, se incrementa. Finalmente, se comprueba la colinealidad de las variables cuantitativas nota2s, edad y nota1s, este análisis se hará en términos de regresión lineal, para descartar correlaciones fuertes entre las variables cualitativas que forman parte del modelo.

Tabla 16

Coeficientes de colinealidad.

Modelo		Estadísticos de colinealidad	
		Tolerancia	FIV
1	Nota1s	,200	5,007
	Nota2s	,200	5,007
Variable dependiente: edad			

Se analiza la (Tabla XVI), en función de la tolerancia (t) y el factor de inflación de la varianza (FIV). Para valores de tolerancia menores que 0,1 (t >0.1), existen problemas de

Tabla 17

Diagnostico de la colinealidad.

Diagnósticos de colinealidad						
Modelo	Dimensión	Autovalores	Índice de condición	Proporciones de la varianza		
				(Constante)	Nota1s	Nota2s
1	1	2,883	1,000	,02	,00	,00
	2	,101	5,333	,89	,03	,08
	3	,016	13,529	,09	,97	,92

correlación entre las variables, entonces sería necesario eliminar una de las variables. Por otro lado, el FIV debe ser mayor a 10 (FIV>10), este criterio varía según los autores.

En el caso de estudio, tomando como variable dependiente a edad (Tabla XVI), el valor de t = 0,2 que es mayor que 0,1 y el valor del FIV = 5,007; se encuentra dentro del rango establecido.

De igual manera se analiza las variables nota2s y edad en relación a nota1s, además nota2s en función de la edad y a nota1s (Tabla XVII).

Los valores de las varianzas en todos los casos son pequeños por lo cual se puede afirmar que no existe colinealidad.

3.2.3. Independencia de errores (no autocorrelación)

Si se incumple este supuesto se dice que existe una auto correlación, por tanto, los coeficientes estadísticos quedaran inválidos, es decir la varianza de los coeficientes no es mínima (errores estándar), es decir que no exista correlación entre el error y las variables explicativas. Para comprobar este supuesto se utiliza la prueba de Durbin-Watson con las variables cuantitativas seleccionadas por el modelo (Nota2s, Edad y Nota1s).

Tabla 18

Test Durbin-Watson.

Modelo	Durbin-Watson
1	1,816

Se analiza el coeficiente en función de la siguiente regla: Si el coeficiente se encuentra entre -1,5 y 2,5, que representa al valor de independencia.

Se debe tener en cuenta que 2 es el valor de independencia perfecta. Para el caso de estudio se tiene 1,816 (Tabla XVIII), por lo cual se acepta la hipótesis que manifiesta que

no existe autocorrelación entre las variables, además se concluye que los estadísticos obtenidos son válidos, es decir que no existe autocorrelación entre las variables lo que comprueba el supuesto.

3.3. Resultados obtenidos

Los datos utilizados para generar el modelo predictivo, lo denominaremos conjunto de datos de entrenamiento. Para verificar la exactitud en la predicción del modelo se tomará un conjunto de datos de prueba.

La (Tabla XIX), muestra en su columna “DESERTOR”, la clasificación del conjunto de datos original, la columna “Probabilidad de deserción” que tiene un estudiante de desertar con determinadas características, la columna “Categoría”, es la clasificación de la categoría según el modelo en función de las variables seleccionadas.

Por ejemplo, el primer estudiante que se muestra en la (Tabla XIX), tiene una probabilidad de desertar del 9,36%, el modelo lo ha clasificado como no desertor que para el caso es igual a los datos originales, sin embargo, el estudiante Nº6, tiene una probabilidad de desertar de 4,61%, el modelo lo ha clasificado como No desertor esta categoría no coincide con el conjunto de datos original, donde se muestra al estudiante como desertor.

La Figura 2, muestra la clasificación del modelo en relación al paso 5, para los grupos observados y probabilidades pronosticadas.

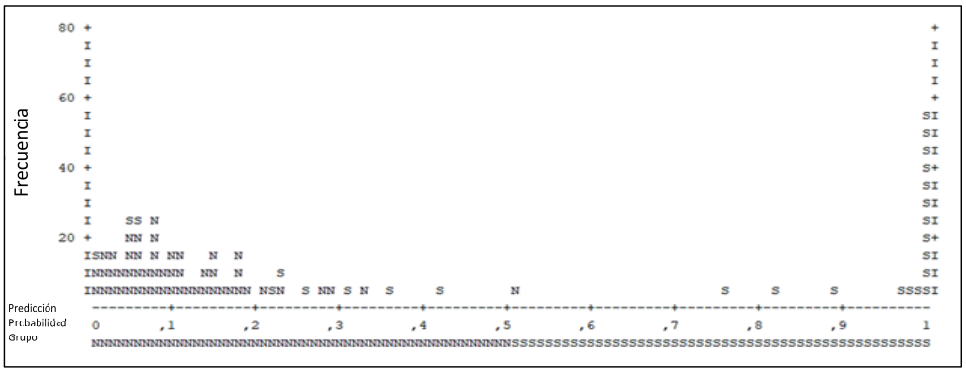

Figura 2

Grupos observados y probabilidades pronosticadas. N: No Desertor, S: Desertor, cada símbolo representa cinco casos.

DOI 10.18502/espoch.v3i1.14477

Tabla 19

Resultados -prueba de modelo.

N°	G	EC	ED	RA	LN	CR	Nota1p	Nota2p	Nota1s	Nota 2s	DESERTOR	Probabilidad de deserción	Categoría
1	MASCULINO	SOLTERO	23	INDIGENA	AMBATO	OTROS	6,6	6,7	7,2	7,4	No Desertor	0,09366	No Desertor
2	FEMENINO	SOLTERO	18	MESTIZA	OTROS	OTROS	8,8	8,1	9,1	8,5	No Desertor	0,03884	No Desertor
3	MASCULINO	SOLTERO	20	MESTIZA	OTROS	OTROS	8,1	8,5	7,7	7,8	No Desertor	0,13942	No Desertor
4	FEMENINO	SOLTERO	20	MESTIZA	OTROS	AMBATO	7	7	7	7	No Desertor	0,14071	No Desertor
5	FEMENINO	CASADO	20	MESTIZA	OTROS	AMBATO	8,6	7,2	7	0	Desertor	0,99398	Desertor
6	MASCULINO	CASADO	22	MESTIZA	OTROS	AMBATO	8	8,1	7,4	7,8	Desertor	0,46115	No Desertor
7	FEMENINO	SOLTERO	21	MESTIZA	OTROS	AMBATO	7	7	7	7	No Desertor	0,10611	No Desertor
8	FEMENINO	SOLTERO	22	MESTIZA	OTROS	AMBATO	8,4	8,7	8,2	8	No Desertor	0,02373	No Desertor
9	MASCULINO	SOLTERO	32	MESTIZA	OTROS	AMBATO	7,5	8,5	9,3	8,8	No Desertor	0,00079	No Desertor
10	MASCULINO	SOLTERO	20	MESTIZA	OTROS	AMBATO	7,8	7,9	8	7,8	No Desertor	0,12278	No Desertor

4. Conclusiones

Las variables significativas que ayudan a la predicción de la deserción estudiantil son: nota2s, estado civil, edad, género y nota1s, en tanto que las menos significativas son: nota1p, nota2p, lugar de nacimiento, ciudad de residencia y raza. En función del parámetro Exp (B) se determinó que existe mayor riesgo de deserción si el estado civil del estudiante es casado y menor riesgo si es divorciado o soltero, además, los estudiantes de género masculino tienen 0,417 más riesgo de desertar respecto a los estudiantes de género femenino.

Dado que la información utilizada para la creación del modelo representa una muestra pequeña es necesario incluir más datos para generar modelos más robustos. El modelo solamente toma en cuenta a cinco de las 10 variables analizadas originalmente, lo que podría significar un defecto. En escenarios como estos se requiere el incremento de variables explicativas.

El modelo desarrollado presento síntomas de multicolinealidad pudiendo representar un problema desde el punto de vista estadístico, por lo que es necesario que se incremente el tamaño de la muestra. De un total de 425 estudiantes, 371 han sido clasificados de manera correcta, esto significa que el 87,3% de la población ha sido clasificada como No desertor o desertor.

References

[1] Durán J, Díaz G. Análisis de la deserción estudiantil en la universidad autónoma metropolitana. Revista iberoamericana de educación superior. 1990;19(2):95–128.

[2] Ovares R. Análisis de las estrategias para la prevención de la deserción y retención de la población estudiantil que lleva a cabo el personal docente y administrativo del Liceo de Miramar. Gestión Education. 2012;2:1–27.

[3] Argote I, Jimenez R, Gómez J. Cuarta Conferencia Latinoamericana sobre el abandono en la Educación Superior. In: Detección de patrones de deserción en los programas de pregrado de la Universidad Mariana de San Juan de Pasto, aplicando el proceso de descubrimiento de conomimientosobre base de datos (KDD) y su implementación en modelos matemáticos de predicción. Colombia; 2014; 1–7.

[4] Girón Cruz LE, González Gómez DE. Determinantes del rendimiento académico y la deserción estudiantil, en el programa de Economía de la Pontificia Universidad Javeriana de Cali. EcoGestDesarro. 2005;3:173–201.

[5] Azoumana K. Análisis de la deserción estudiantil en la Universidad Simón Bolívar, Facultad Ingeniería de Sistemas, con Técnicas de minería de datos. Pensam Am. 2013;6(10):41–51.

[6] Más-Estellés. Alcover-Arándiga, Dapena-Janeiro, Valderruten-Vidal, Satorre-Cuerda, Llopis-Pascual, et al. Rendimiento académico de los estudios de Informática en algunos centros españoles. XV Jenui; 2009: 8.

[7] Páramo GJ, Maya CA. Deserción estudiantil universitaria. Conceptualización. Volume 35. Revista Universidad EAFIT; 2012; 65–78.

[8] Bernardo Gutiérrez AB, Cerezo Menéndez R, Rodríguez-Muñiz LJ, Núñez Pérez JC, Tuero Herrero E, Esteban García M. Predicción del abandono universitario: variables explicativas y medidas de prevención. Rev. Rafael Rodriguez-Fuentes [Internet]. 2015;(16):63–84. Available from: https://revistascientificas.us.es/index.php/fuentes/article/view/2363

[9] González MT. Absentismo y abandono escolar: una situación singular de la exclusión educativa. Revista Iberoamericana sobre Calidad., Eficacia y Cambio en Educación. 2006;4(1):1–15.

[10] Sifuentes Bitocchi O. Modelos predictivos de la deserción estudiantil en una universidad privada peruana. inD Dataset. 2018;21(2):47.

[11] Barrero J, Garzón G, Gómez Ó. Variables asociadas con el fenómeno de la deserción de los estudiantes en la Fundación Universitaria Konrad Lorenz. Pensando Psicología. 2013;9(16):55–68.

[12] Martínez AF, Márquez JC, Martín BC, Alonso SS, Campos JC. Predicción de lesiones deportivas mediante modelos matemáticos. Apunts. Medicina de l'Esport. 2008;43(157):41–44.

[13] Balaguer P. Una explicación del rendimiento estudiantil universitario mediante modelos de regresión logística. Visión Gerencial. 2009;0(2):415–427.

[14] Bonaldo L, Pereira LN. Dropout: Demographic Profile of Brazilian University Students [Internet]. Procedia - Social and Behavioral Sciences. 2016;228(June):138–143.

[15] García MV, Alvarado J, Jiménez A. La predicción del rendimiento académico: regresión lineal versus regresión logística. Psicothema. 2000;12:222–248.

[16] Ponsot E, Sinha S, Varela L, Varela J. Un modelo de regresión logística del rendimiento en los estudios universitarios: Caso FACES-ULA. 2009.

[17] Reyes Rocabado J, Escobar Flores C, Duarte Vargas J, Ramírez Peradotto P. Una aplicacion del modelo de regression logistica en la prediccion del rendimiento estudiantil. Estud Pedagóg (Valdivia). 2007;33(2):101–120.

[18] Santosa RG, Chrismanto AR. Logistic Regression model for predicting first semester GPA category based on high school academic achievement. Researchers World : Journal of Arts Science & Commerce. 2017;VIII(2): Researchers World : Journal of Arts Science & Commerce 1–12.

[19] Lichtenberger E, George-Jackson C. Predicting high school students' interest in majoring in a STEM field: Insight into high school students' post-secondary plans. Journal of Career and Technical Education. 2012;28(1). https://doi.org/10.21061/jcte.v28i1.571.

[20] Constitución de la Republica del Ecuador. Publicada en el Registro Oficial 449 de 20 de octubre de 2008 [Internet]. Registro Oficial 449 20 October 2008. Quito Ecuador; 2008. Available from: https://www.oas.org/juridico/pdfs/mesicic4_ecu_const.pdf

[21] Calderón MG, Espinel EE, Garzón PV, Pástor CR. Impacto social de la deserción estudiantil en la Facultad de Ciencias Químicas en primer semestre de la Universidad Central del Ecuador. Polo del Conocimiento. 2017;2(8):65.

[22] Castro B, Rivas G. Estudio sobre el fenómeno de la deserción y retención escolar en localidades de alto riesgo. Soc Hoy. 2006;(11):35–72.

[23] Silvente VB, Baños RV. Cómo obtener un Modelo de Regresión Logística Binaria con SPSSRevista Innova Educación. 2014;8(2):105–118.

[24] López-Roldán P, Fachelli S. Metodología de la investigación social cuatitativa [Internet]. Primera Ed. Universidad Autónoma de Barcelona, editor. Universidad Autónoma de Barcelona. Barcelona; 2016:55. Available from: https://ddd.uab.cat/pub/caplli/2016/163570/metinvsoccua_a2016_cap3-10.pdf

[25] Hosmer D, Stanley L. Regresión Logístic Aplicada. Segunda Ed. Vol. 70, Journal of Environmental Health. Ohio: Wiley series de probabilidad y estadística; 2007.

[26] Clifford B, Taylor R. Bioestadística. Primera Ed. PEARSON, editor. Bioestadística. México; 2014:531.

[27] Llaugel FA, Fernández AI. Evaluación del uso de modelos de regresión logística para el diagnóstico de Instituciones Financieras. Revista Internacional de Investigación en Ciencias Sociales. 2011;36(4):XXXVI.

[28] Field A, Miles J. Discovering Statistics using R [Internet]. Primera Ed. Vol. 1. Londres: SAGA Publications; 2012:957. Available from: https://nyu-cdsc.github.io/learningr/assets/discoveringstatistics.pdf