INGENIUM ET POTENTIA
Revista Electrónica Multidisciplinaria de Ciencias Básicas, Ingeniería y Arquitectura
Año VIII. Vol VIII. N°14. Enero-Junio. 2026
Hecho el depósito de Ley: FA2019000052
FUNDACIÓN KOINONIA (F.K).
Santa Ana de Coro, Venezuela.

https://doi.org/10.35381/i.p.v8i14.5061

Modelado predictivo de la deserción estudiantil en el Instituto Superior Tecnológico General Eloy Alfaro, Ecuador

Predictive Modeling of Student Dropout Rates at the General Eloy Alfaro Higher Technological Institute, Ecuador

Ana María Pilco-Salazar
anam.pilco@unach.edu.ec
Universidad Nacional de Chimborazo; Riobamba, Chimborazo
Ecuador
https://orcid.org/0000-0003-2380-037X

Sayuri Monserrath Bonilla-Novillo
sayuri.bonilla@unach.edu.ec
Universidad Nacional de Chimborazo; Riobamba, Chimborazo
Ecuador
https://orcid.org/0000-0001-6509-8238

Recibido: 20 de enero 2026
Revisado: 21 de marzo 2026
Aprobado: 15 de abril 2026
Publicado: 01 de mayo 2026

INGENIUM ET POTENTIA
Revista Electrónica Multidisciplinaria de Ciencias Básicas, Ingeniería y Arquitectura
Año VIII. Vol VIII. N°14. Enero-Junio. 2026
Hecho el depósito de Ley: FA2019000052
FUNDACIÓN KOINONIA (F.K).
Santa Ana de Coro, Venezuela.

RESUMEN

La deserción estudiantil afecta la permanencia y culminación de los estudios en la educación superior. Este estudio tuvo como objetivo desarrollar un modelo predictivo de deserción estudiantil mediante regresión logística y análisis multivariable en el Instituto Superior Tecnológico General Eloy Alfaro. La investigación tuvo enfoque cuantitativo, diseño no experimental y carácter retrospectivo, con registros institucionales de 684 estudiantes de los períodos 2021-I a 2025-I. Se analizaron variables académicas, sociodemográficas, socioeconómicas e institucionales mediante análisis bivariado, evaluación de multicolinealidad y regresión logística binaria. Los resultados mostraron que las variables académicas tuvieron relación importante con la deserción en el análisis inicial; sin embargo, luego de la depuración por VIF, el modelo final quedó conformado por género, ocupación del estudiante y estado civil agrupado. El modelo obtuvo una exactitud de 0,665 y un AUC de 0,6623. Se concluye que la regresión logística puede apoyar el seguimiento institucional de estudiantes en riesgo.

Descriptores: Deserción escolar; enseñanza superior; estudiante universitario; análisis cuantitativo; estudio social. (Tesauro UNESCO)

ABSTRACT

Student dropout rates affect retention and completion rates in higher education. The objective of this study was to develop a predictive model of student dropout using logistic regression and multivariate analysis at the General Eloy Alfaro Higher Technological Institute. The research employed a quantitative approach, a non-experimental design, and a retrospective methodology, utilizing institutional records of 684 students from the 2021-I to 2025-I periods. Academic, sociodemographic, socioeconomic, and institutional variables were analyzed using bivariate analysis, multicollinearity assessment, and binary logistic regression. The results showed that academic variables were significantly associated with dropout in the initial analysis; however, after VIF-based variable screening, the final model consisted of gender, student occupation, and grouped marital status. The model achieved an accuracy of 0.665 and an AUC of 0.6623. It is concluded that logistic regression can support institutional monitoring of at-risk students.

Descriptors: School dropout; higher education; university student; quantitative analysis; social study. (UNESCO Thesaurus)

INTRODUCCIÓN

La deserción estudiantil en la educación superior representa uno de los desafíos más significativos en los sistemas educativos vigentes, debido a su gran alcance, este fenómeno no solo afecta la eficiencia institucional de educación superior, esto pone en evidencia la imparcialidad en el acceso al conocimiento, el desarrollo socioeconómico y evidencia la no permanencia de los estudiantes en el sistema educativo (Quincho Apumayta, y otros, 2025). Este acontecimiento se define como la interrupción, definitiva o temporal, de los estudios antes de la obtención del título académico, y se distingue por la naturaleza de múltiples factores, en la que influyen variables académicas, socioeconómicas, demográficas y contextuales (Rodríguez, Espinoza, Ramírez, & Ganga, 2018). En el ámbito latinoamericano, la deserción presenta niveles alarmantes, particularmente en instituciones tecnológicas, donde las situaciones económicas, la integración laboral temprana y las dificultades académicas contribuyen a incrementar la inestabilidad de los estudiantes (Villegas, Núñez, & Luis, 2024).

Desde un marco de referencia académica, la literatura reciente ha demostrado que el desempeño académico conforma uno de los principales aspectos decisivos de la continuidad del sistema educativo, así como la innovación es esencial en las prácticas pedagógicas, descartando las clásicas metodologías tradicionales, logrando así un fuerte impacto en el rendimiento académico. Particularmente, el bajo rendimiento en los primeros niveles de educación se demuestra en la no aprobación de asignaturas y la acumulación de materias pendientes, lo que genera retrasos en el avance académico de los estudiantes que se vincula directamente como una mayor amenaza de deserción, estos factores afectan la continuidad y la permanencia de los estudiantes en el sistema educativo (Villegas, Núñez, & Luis, 2024). Este desempeño puede atribuirse, en gran medida, a las dificultades que afrontan los estudiantes para adaptarse al entorno de educación superior, el reducido fortalecimiento de conocimientos fundamentales que limitan la capacidad de los estudiantes para comprender y relacionar contenidos más

avanzados, tomando en cuenta también la falta de estrategias de aprendizaje autodidacta (Morales, Xicoténcatl Ramírez, Ibarra Corona, & García Reyes, 2024). En este sentido, la reprobación de asignaturas no solo refleja un bajo rendimiento académico, sino que también actúa como un indicador temprano de riesgo dentro de las trayectorias educativas, la acumulación de módulos sin aprobación puede generar efectos negativos, como la desmotivación, factores que en conjunto aumenta el rezago académico, los cuales nos permite identificar estudiantes susceptibles de abandonar sus estudios (Gutiérrez Monsalve, Garzón, & Segura Cardona, 2021).

Adicionalmente, diversos estudios han señalado que factores sociodemográficos, como la edad, el estado civil y la situación laboral, influyen en la deserción estudiantil, aunque con menor intensidad en comparación con los factores académicos, su nivel de compromiso con la formación y sus prioridades personales, especialmente en etapas que se asumen mayores obligaciones familiares o laborales (Cárdenas Matute, Valle Franco, & Tapia Segarra, 2023). En contextos donde los estudiantes combinan actividades laborales con la formación académica, la carga de responsabilidades tiende a incidir de manera significativa en el rendimiento y la continuidad de los estudios; esta evidencia da una clave para considerar que estos elementos pueden aumentar la vulnerabilidad frente a la deserción del sistema educativo; este efecto suele ser indirecto y estimado por el desempeño académico, la integración institucional y el acceso a recursos de apoyo. (Villegas, Núñez, & Luis, 2024). Sin embargo, estos factores suelen interactuar con variables académicas, configurando un sistema complejo de relaciones que requiere ser analizado mediante enfoques multivariables, ya que la necesidad de distribuir el tiempo que dispone el estudiante, entre el trabajo, las tareas personales y las exigencias académicas suelen limitar las horas disponibles para el estudio, así como también la asistencia regular a clases y la participación en diversas actividades formativas complementarias (Carvajal, González, & Sarzoza, 2018).

En este contexto, el desarrollo de la matemática computacional y el análisis de datos ha permitido abordar el problema de la deserción desde una perspectiva cuantitativa, mediante la construcción de modelos predictivos capaces de identificar patrones y estimar probabilidades de abandono (Aulck, Velagapudi, Blumenstock, & West, 2016). Entre las técnicas más utilizadas, la regresión logística se ha consolidado como una herramienta fundamental para el análisis de variables binarias, debido a su capacidad para modelar la probabilidad de ocurrencia de un evento en función de múltiples variables explicativas (Alcaraz González, Morales Benítez, González García, & Paredes Medina, 2024). Este modelo se basa en la transformación logit, que establece una relación lineal entre los predictores y el logaritmo de las probabilidades, facilitando su interpretación a través de los odds ratio (Hosmer Jr, Lemeshow, & Sturdivant, 2013).

A pesar del avance de técnicas más complejas basadas en aprendizaje automático, como redes neuronales y árboles de decisión, estudios destacan que la regresión logística continúa siendo ampliamente utilizada en el ámbito educativo, debido a su interpretabilidad, estabilidad en la estimación de parámetros y menor complejidad computacional, estas características la convierten en una herramienta adecuada para estudios aplicados, donde la interpretación de los resultados es tan relevante como la capacidad predictiva del modelo (Aguilar Reyes, Mejía Peñafiel, Morocho Barrionuevo, & Velasco Castelo, 2025).

No obstante, la construcción de modelos predictivos confiables requiere la aplicación de procedimientos estadísticos complementarios que garanticen la validez de las estimaciones (Fernández Casal, Costa Bouzas, & Oviedo de la Fuente, 2024). En este sentido, el análisis de correlación permite identificar relaciones lineales entre variables, mientras que la detección de multicolinealidad mediante el factor de inflación de la varianza (VIF) resulta fundamental para evitar la redundancia de información en la matriz de diseño (Salmerón, García, & García, 2020). La presencia de multicolinealidad puede generar inestabilidad en los coeficientes del modelo y afectar su capacidad explicativa,

por lo que su identificación y corrección constituyen una etapa crítica en el proceso de modelado (Salmerón Gómez & Rodríguez Martínez, 2017).

Por otra parte, el desarrollo del learning analytics ha impulsado la integración de datos académicos, socioeconómicos y comportamentales en el análisis de la deserción estudiantil. Estos enfoques permiten no solo predecir el abandono, sino también diseñar estrategias de intervención temprana orientadas a mejorar la retención. En particular, la identificación de estudiantes en riesgo mediante modelos predictivos facilita la implementación de políticas institucionales basadas en evidencia, contribuyendo a la mejora de los indicadores de permanencia estudiantil. (Abdulkareem Shafiq, Marjani, Ariyaluran Habeeb, & Asirvatham, 2022).

A pesar de estos avances, en el contexto ecuatoriano, especialmente en instituciones tecnológicas, existe una limitada aplicación sistemática de modelos estadísticos multivariables que integren técnicas de regresión logística, análisis de correlación y evaluación de multicolinealidad. (Nigri, Bilancia, & Cafarelli, 2025). Esta situación evidencia una brecha en la literatura científica reciente, así como una oportunidad para el desarrollo de investigaciones que contribuyan al fortalecimiento de la toma de decisiones en el ámbito educativo (Morales, Xicoténcatl Ramírez, Ibarra Corona, & García Reyes, 2024).

En este sentido, el presente estudio tiene como objetivo desarrollar un modelo predictivo de la deserción estudiantil mediante regresión logística y análisis multivariable, utilizando datos del Instituto Superior Tecnológico General Eloy Alfaro. Se plantea como hipótesis que las variables académicas presentan una relación importante con la deserción estudiantil; sin embargo, su permanencia en el modelo final dependerá de los procesos de depuración estadística, especialmente del análisis de multicolinealidad. A partir de este enfoque, se busca estimar la probabilidad de abandono e identificar los factores que aportan información relevante para orientar estrategias institucionales de seguimiento y permanencia estudiantil.

MÉTODO

La investigación se realizó bajo un enfoque cuantitativo, que permitió analizar la problemática a partir de datos numéricos, con la finalidad de medir, comparar y asociar variables para obtener los resultados. La investigación no fue experimental debido a que las variables vinculadas a la deserción estudiantil no han sido manipuladas, no se tuvo intervención directa o indirecta sobre los estudiantes, ya sea sobre sus condiciones socioeconómicas, académicas o étnicas. El alcance de la investigación es explicativo debido a que se identificaron y analizaron varios factores que influían en la deserción estudiantil mediante el modelado predictivo. Para la elaboración del código en Python se siguió un diagrama de flujo (Figura 1).

La información analizada proviene del Instituto Superior Tecnológico General Eloy Alfaro, el cual se encuentra ubicado en la provincia de Orellana, en el cantón La Joya de los Sachas. Sus funciones toman inicio a partir de 2020 en la región Amazónica, por lo cual es clave el análisis de la deserción estudiantil en sus primeros semestres; por ello, el periodo que se analizó incluye datos históricos desde 2021-I a 2025-I.

El instituto cuenta actualmente con tres carreras, Tecnología Superior en Administración, Tecnología Superior en Mecánica Industrial y Tecnología Superior en Educación Inicial, estas carreras tienen diferentes características académicas, razón por la cual se las tomaron para la investigación. La matriz de análisis contiene información académica y estudiantil consolidada a partir de los registros del Sistema Integrado de Gestión Académica (SIGA) y del Sistema Integral de Información de la Educación Superior (SIIES), administrado por la SENESCYT. La población de estudio está conformada por un total de 684 estudiantes, los cuales pertenecían a las tres carreras ofertadas por el instituto.

La matriz analizada estuvo conformada por 27 variables entre las que se consideró información demográfica como género, edad, estado civil y etnia; variables socioeconómicas como el bono de desarrollo, beca, ocupación, ingreso total del hogar y

número total de miembros en el hogar; variables académicas como si el estudiante ha repetido materias, asignaturas perdidas, porcentaje de asistencia y promedio; así como variables institucionales que están ligadas con la carrera y la modalidad con la que se conllevo las clases. Estas variables permitieron desarrollar un modelado predictivo y análisis multivariable de la deserción estudiantil en el Instituto Superior Tecnológico Eloy Alfaro (Tabla 1).

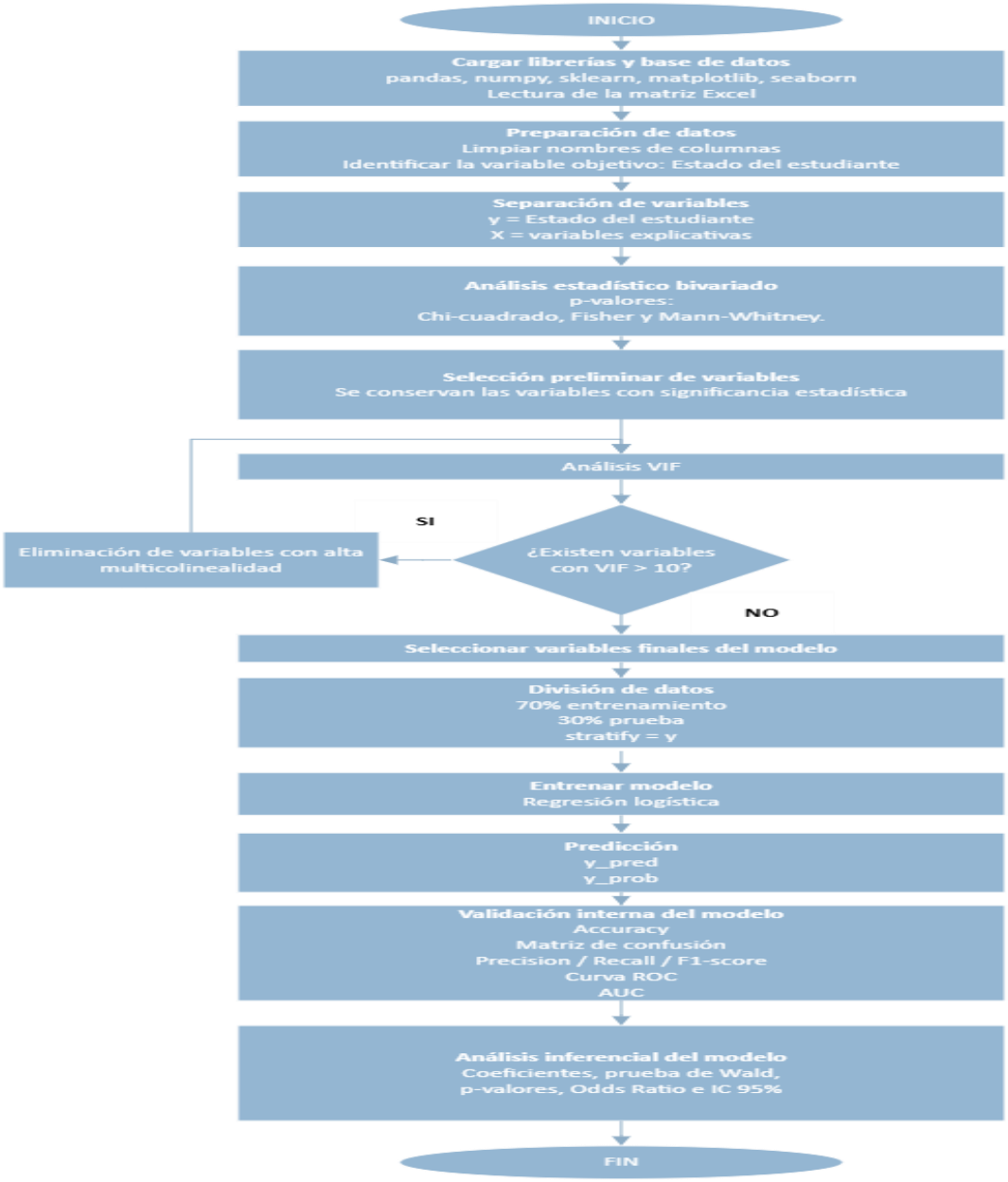

Figura 1. Diagrama de flujo del análisis y modelado predictivo de la deserción estudiantil.
Elaboración: Los autores.

Tabla 1.
Variables de la matriz analizada.

N.º	Variable	Tipo de información
1	ID	Identificación del estudiante
2	Género	Información sociodemográfica
3	Estado civil	Información sociodemográfica
4	Etnia	Información sociodemográfica
5	Discapacidad	Información sociodemográfica
6	Fecha de Nacimiento	Información sociodemográfica
7	Edad	Información sociodemográfica
8	Provincia de Nacimiento	Información geográfica
9	Cantón de Nacimiento	Información geográfica
10	Provincia de Residencia	Información geográfica
11	Cantón de Residencia	Información geográfica
12	Tipo de Colegio	Información académica previa
13	Modalidad de Carrera	Información académica
14	Ha repetido al menos una materia	Información académica
15	Ocupación del estudiante	Información socioeconómica
16	Bono de desarrollo	Información socioeconómica
17	Recibe Beca	Información socioeconómica
18	Nivel de formación del padre	Información familiar
19	Nivel de formación de la madre	Información familiar
20	Ingresos totales del hogar	Información socioeconómica
21	Cantidad de miembros en el hogar	Información familiar
22	Estado del estudiante	Variable objetivo
23	Nivel de deserción	Información académica asociada a la deserción
24	Asignaturas perdidas	Información académica
25	Porcentaje de asistencia	Información académica
26	Promedio	Información académica
27	Carrera	Información académica

Elaboración: Los autores.

Tabla 2.
Calculó el vector de corrección.

Componente de (g)	Variable	Casos con (X=1)	Desertores (Y=1)	No desertores (Y=0)	Cálculo aplicado	Resultado
g0	Intercepto	478	210	268	$((210 \times 0.5) + (268 \times -0.5))$	-29
g1	Género_2	333	123	210	$((123 \times 0.5) + (210 \times -0.5))$	-43.5
g2	Ocupación del estudiante_1	189	63	126	$((63 \times 0.5) + (126 \times -0.5))$	-31.5
g3	Estado civil agrupado_Estado civil 2	56	26	30	$((26 \times 0.5) + (30 \times -0.5))$	-2
g4	Estado civil agrupado_Otros	27	19	8	$((19 \times 0.5) + (8 \times -0.5))$	5.5

Elaboración: Los autores.

Con los datos de entrenamiento se obtuvo:

$$g = \begin{bmatrix} -29 \\ -43.5 \\ -31.5 \\ -2 \\ 5.5 \end{bmatrix}$$

Este resultado indica el ajuste que deben seguir los coeficientes para mejorar la estimación del modelo. Luego se calculó la matriz de información:

$$H = X^T W X$$

Donde:

- H: matriz de información del modelo. Sirve para actualizar los coeficientes β.
- X: matriz de predictores. Contiene la columna del intercepto y las variables X_1, X_2, X_3, X_4.
- X^T: matriz transpuesta de X. Permite operar matemáticamente con los errores y las variables.
- W: matriz de pesos, corresponde al término de regularización usado para estabilizar el cálculo. En la primera iteración, como $p = 0.5$, se tiene:

$$W = p(1-p)$$

$$W = 0.5(1-0.5) = 0.25$$

En la Tabla 3 se muestra el cálculo de la matriz de información.

Tabla 3.
Cálculo de la matriz de información $H = X^T WX$

Elemento	Variables relacionadas	Coincidencias	Cálculo (X^TWX)	Valor final
(H_00)	Intercepto–Intercepto	478	(478(0.25)=119.50)	119.50
(H_01)	Intercepto–Género_2	333	(333(0.25)=83.25)	83.25
(H_02)	Intercepto–Ocupación_1	189	(189(0.25)=47.25)	47.25
(H_03)	Intercepto–Estado civil 2	56	(56(0.25)=14.00)	14.00
(H_04)	Intercepto–Otros	27	(27(0.25)=6.75)	6.75
(H_11)	Género_2–Género_2	333	(333(0.25)=83.25)	83.25
(H_12)	Género_2–Ocupación_1	157	(157(0.25)=39.25)	39.25
(H_13)	Género_2–Estado civil 2	47	(47(0.25)=11.75)	11.75
(H_14)	Género_2–Otros	18	(18(0.25)=4.50)	4.50
(H_22)	Ocupación_1–Ocupación_1	189	(189(0.25)=47.25)	47.25
(H_23)	Ocupación_1–Estado civil 2	26	(26(0.25)=6.50)	6.50
(H_24)	Ocupación_1–Otros	3	(3(0.25)=0.75)	0.75
(H_33)	Estado civil 2–Estado civil 2	56	(56(0.25)=14.00)	14.00
(H_34)	Estado civil 2–Otros	0	(0(0.25)=0.00)	0.00
(H_44)	Otros–Otros	27	(27(0.25)=6.75)	6.75

Elaboración: Los autores.

Como la matriz es simétrica, los valores del lado inferior se repiten. Por tanto:

$$H = \begin{bmatrix} 119.50 & 83.25 & 47.25 & 14.00 & 6.75 \\ 83.25 & 83.25 & 39.25 & 11.75 & 4.50 \\ 47.25 & 39.25 & 47.25 & 6.50 & 0.75 \\ 14.00 & 11.75 & 6.50 & 14.00 & 0.00 \\ 6.75 & 4.50 & 0.75 & 0.00 & 6.75 \end{bmatrix}$$

Con estos valores, la primera actualización de los coeficientes se realizó mediante:

$$\beta^{(1)} = \beta^{(0)} + H^{-1}g$$

Al resolver esta operación se obtuvo la primera aproximación:

$$\beta^{(1)} = \begin{bmatrix} 0.4077 \\ -0.7927 \\ -0.4620 \\ 0.3073 \\ 0.8596 \end{bmatrix}$$

El proceso se repitió hasta que los coeficientes presentaron cambios mínimos entre una iteración y otra. La convergencia del modelo se resume en la Tabla 4:

Tabla 4.
Iteraciones para la estimación de los coeficientes del modelo logístico.

Iteración	(\beta_0)	(\beta_1) Género_2	(\beta_2) Ocupación_1	(\beta_3) Estado civil 2	(\beta_4) Otros
Inicial	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.4520	-0.9963	-0.3200	0.3953	0.9854
2	0.4735	-1.0429	-0.3506	0.4304	1.0803
3	0.4737	-1.0434	-0.3510	0.4307	1.0817
4	0.4737	-1.0434	-0.3510	0.4307	1.0818

Elaboración: Los autores.

Con base en este procedimiento, el modelo logístico estimado quedó expresado como:

$$Z = 0.4737 - 1.0434X1 - 0.3510X2 + 0.4307X3 + 1.0817X4$$

De esta manera, el cálculo matemático permitió evidenciar que los coeficientes del modelo se obtuvieron mediante un proceso iterativo de estimación, y que los resultados generados en Python corresponden a la aplicación automatizada del mismo procedimiento matemático.

Para hacer una validación interna, en Python se dividió la base de datos en dos partes: una para entrenar el modelo y otra para probarlo. El 70 % de los registros se usó para el entrenamiento, y el 30 % restante para revisar qué tan bien predecía. Este conjunto de

prueba sirvió como una validación exploratoria, ayudando a comprobar cómo se comportaba el modelo con datos que no había visto durante el entrenamiento.

Para medir el desempeño del modelo se utilizó la matriz de confusión, conformada por verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP)y falsos negativos (FN).

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN}$$

$$Precisión = \frac{VP}{VP + FP}$$

$$Recall = \frac{VP}{VP + FN}$$

$$F1\text{-}score = 2\left(\frac{Precisión \times Recall}{Precisión + Recall}\right)$$

La curva ROC se construyó a partir de la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos:

$$TPR = \frac{VP}{VP + FN}$$

$$FPR = \frac{FP}{FP + VN}$$

El área bajo la curva (AUC)permitió evaluar la capacidad del modelo para diferenciar entre estudiantes desertores y no desertores.

De forma complementaria, se estimaron los coeficientes del modelo logístico y su significancia estadística. La prueba de Wald se calculó mediante:

$$Wald = \left(\frac{\beta}{EE}\right)^2$$

donde βrepresenta el coeficiente estimado y EEel error estándar. El Odds Ratio se obtuvo con:

$$OR = e^{\beta}$$

Finalmente, los intervalos de confianza al 95 % para los Odds Ratio se calcularon como:

$$IC_{95\%} = e^{\beta \pm 1.96(EE)}$$

Estas métricas permitieron evaluar tanto el desempeño predictivo del modelo como el aporte individual de cada variable incluida en la regresión logística.

En resumen, la metodología que se aplicó permitió llevar un proceso bien organizado de limpieza, preparación, análisis y modelado de los datos, todo con el objetivo de reconocer los factores que más explicaban la deserción estudiantil. La mezcla de técnicas como la depuración de variables, la revisión de la multicolinealidad y el uso de la regresión logística binaria facilitaron armar un modelo predictivo basado tanto en criterios estadísticos sólidos como en una interpretación coherente.

RESULTADOS

El análisis estadístico bivariado permitió reconocer las variables que presentaron asociación con el estado del estudiante. Para ello, se aplicaron las pruebas de Chi-cuadrado, Fisher y Mann-Whitney, conforme al tipo de variable analizada.

Los resultados señalaron que las variables académicas tuvieron una mayor conexión con la deserción estudiantil. Entre ellas destacaron el promedio académico, la repetición de al menos una materia, las asignaturas perdidas y el porcentaje de asistencia. Esto indica que el desempeño académico del estudiante tiene un papel importante dentro del fenómeno de la deserción.

También se encontraron asociaciones significativas en algunas variables sociodemográficas y socioeconómicas, como el género, estado civil, ocupación del

estudiante, lugar de nacimiento y lugar de residencia. Sin embargo, estas variables mostraron una relación menor en comparativa con las variables académicas.

Estos resultados sirvieron para hacer una primera depuración de variables para las siguientes fases del análisis. En la Tabla 5 se presentan los valores obtenidos en el análisis estadístico bivariado.

Tabla 5.

a. Resultados del análisis estadístico bivariado.

Variable	Tipo de variable	Prueba aplicada	Estadístico	p-valor	Interpretación
Género	Categórica / codificada	Chi-cuadrado	39,381	0,000	Significativa
Ha repetido al menos una materia	Categórica / codificada	Chi-cuadrado	614,824	0,001	Significativa
Promedio	Numérica	Mann-Whitney U	114943,000	0,001	Significativa
Carrera	Categórica / codificada	Chi-cuadrado	25,125	0,001	Significativa
Asignaturas perdidas	Categórica / codificada	Chi-cuadrado	684,000	0,001	Significativa
Porcentaje de asistencia	Categórica / codificada	Chi-cuadrado	679,949	0,001	Significativa
Ocupación del estudiante	Categórica / codificada	Chi-cuadrado	12,976	0,001	Significativa
Cantón de Nacimiento	Categórica / codificada	Chi-cuadrado	61,991	0,003	Significativa
Modalidad de Carrera	Categórica / codificada	Chi-cuadrado	7,274	0,007	Significativa
Provincia de Nacimiento	Categórica / codificada	Chi-cuadrado	25,151	0,022	Significativa
Estado civil	Categórica / codificada	Chi-cuadrado	10,960	0,027	Significativa

Provincia de Residencia	Categórica / codificada	Chi-cuadrado	14,218	0,047	Significativa
Discapacidad	Categórica / codificada	Chi-cuadrado	3,609	0,058	No significativa
Tipo de Colegio	Categórica / codificada	Chi-cuadrado	7,388	0,061	No significativa

b. Resultados del análisis estadístico bivariado.

Variable	Tipo de variable	Prueba aplicada	Estadístico	p-valor	Interpretación
Etnia	Categórica / codificada	Chi-cuadrado	11,842	0,066	No significativa
Cantón de Residencia	Categórica / codificada	Chi-cuadrado	14,785	0,097	No significativa
Ingresos totales del hogar	Categórica / codificada	Chi-cuadrado	5,589	0,232	No significativa
Edad	Numérica	Mann-Whitney U	55178,000	0,344	No significativa
Fecha de Nacimiento	Categórica / codificada	Chi-cuadrado	640,680	0,408	No significativa
Nivel de formación del padre	Categórica / codificada	Chi-cuadrado	2,603	0,457	No significativa
Bono de desarrollo	Categórica / codificada	Chi-cuadrado	0,229	0,633	No significativa
Recibe Beca	Categórica / codificada	Prueba exacta de Fisher	0,425	0,635	No significativa
Nivel de formación de la madre	Categórica / codificada	Chi-cuadrado	5,385	0,716	No significativa

Cantidad de miembros en el hogar	Numérica	Mann-Whitney U	57126,000	0,851	No significativa

Elaboración: Los autores.

A partir de los resultados del análisis bivariado, se realizó una selección preliminar de variables considerando aquellas que presentaron significancia estadística, es decir, un valor de $p < 0,05$. Este procedimiento permitió reducir el conjunto inicial de variables y conservar las que mostraron mayor relación con el estado del estudiante.

No obstante, esta selección fue considerada como una fase inicial del análisis, ya que después se examinó la posible presencia de multicolinealidad entre las variables elegidas.

Luego del filtro preliminar de variables, se realizó el análisis de multicolinealidad mediante el Factor de Inflación de la Varianza (VIF). En la Tabla 6 se observa el primer análisis realizado con las variables que resultaron significativas en el análisis bivariado.

En esta primera revisión, algunas variables mostraron valores de VIF superiores a 10, lo que indica una alta relación entre ellas. Esta situación se evidenció principalmente en variables académicas como promedio, porcentaje de asistencia y asignaturas perdidas, por lo que fue necesario hacer una depuración.

Tabla 6.
Análisis inicial de multicolinealidad mediante VIF.

Índice	Variable	VIF
0	Promedio	55.976024
16	Porcentaje de asistencia_4	55.790453
18	Modalidad de Carrera_4	35.169993
3	Carrera_2	34.251770
2	Ha repetido al menos una materia_1	19.765668
6	Asignaturas perdidas_7	16.215618

1	Género_2	7.180635
10	Asignaturas perdidas_22	3.660119
9	Asignaturas perdidas_18	3.561252
4	Carrera_3	3.371481
13	Asignaturas perdidas_35	2.954910
5	Asignaturas perdidas_3	2.367142
15	Porcentaje de asistencia_3	2.356636
14	Porcentaje de asistencia_2	2.326477
17	Ocupación del estudiante_1	1.800805
7	Asignaturas perdidas_14	1.737165
8	Asignaturas perdidas_17	1.649977
11	Asignaturas perdidas_25	1.357281
12	Asignaturas perdidas_33	1.236905
19	Estado civil agrupado_Estado civil 2	1.180742
20	Estado civil agrupado_Otros	1.117161

Elaboración: Los autores.

Después de eliminar las variables con VIF > 10, se hizo un segundo análisis. En la Tabla 7 se aprecia que las variables restantes presentaron valores aceptables de VIF, lo que permitió seguir con el proceso de modelado con menor riesgo de multicolinealidad.

Tabla 7.
Análisis de multicolinealidad después de la depuración de variables.

Índice	Variable	VIF
0	Género_2	1.756815
1	Ocupación del estudiante_1	1.578214
2	Estado civil agrupado_Estado civil 2	1.128706
3	Estado civil agrupado_Otros	1.056507

Elaboración: Los autores.

Con las variables finales se construyó el modelo de regresión logística binaria. Para su validación, la base de datos se partió en 70 % para entrenamiento y 30 % para prueba.

En la Figura 2 se muestra la matriz de confusión obtenida. El modelo clasificó correctamente a 93 estudiantes no desertores y 44 estudiantes desertores. También se registraron 23 falsos positivos y 46 falsos negativos, lo que indica que el modelo tuvo un rendimiento moderado en la clasificación del estatus del estudiante.

El modelo alcanzó una exactitud de 0,665. Además, el recall de la clase desertor fue de 0,49, lo que significa que consiguió identificar aproximadamente a la mitad de los estudiantes que realmente desertaron. Este resultado señala que el modelo puede servir como una primera herramienta de apoyo para el análisis de la deserción, aunque no debe tomarse como un mecanismo definitivo de predicción.

A partir de los coeficientes calculados en la regresión logística, se definió el modelo matemático final para estimar la probabilidad de deserción estudiantil. Estos coeficientes fueron verificados mediante el desarrollo manual del procedimiento de estimación, encontrándose concordancia con los resultados obtenidos en Python. Esto confirma que el software aplicó el mismo fundamento matemático de la regresión logística, basado en la función logit y la estimación iterativa.

```
CASO SIN ELIMINACIÓN POR FUGA
MODELO DEPURADO SOLO POR VIF
==================================
Accuracy: 0.665

Matriz de confusión:
[[93 23]
 [46 44]]

Reporte de clasificación:
              precision    recall  f1-score   support

           0       0.67      0.80      0.73       116
           1       0.66      0.49      0.56        90

    accuracy                           0.67       206
   macro avg       0.66      0.65      0.64       206
weighted avg       0.66      0.67      0.66       206

AUC: 0.6623
```

Figura 2. Resultados de evaluación del modelo de regresión logística.
Elaboración: Los autores.

El modelo quedó conformado por cuatro variables predictoras: género, ocupación del estudiante, estado civil 2 y estado civil agrupado en la categoría "Otros". La ecuación lineal del modelo se expresó de la siguiente forma:

$$Z = 0.4264 - 0.8254(\text{Género\_2}) - 0.4961(\text{Ocupación del estudiante\_1}) + 0.3282(\text{Estado civil 2}) + 0.9216(\text{Estado civil Otros})$$

Al aplicar la función logística, la probabilidad estimada de deserción se calculó mediante la siguiente expresión:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-z}}$$

Por tanto:

$$P(\text{deserción}) = \frac{1}{1 + e^{-(0.4264 - 0.8254(\text{Género\_2}) - 0.4961(\text{Ocupación del estudiante\_1}) + 0.3282(\text{Estado civil 2}) + 0.9216(\text{Estado civil Otros}))}}$$

Esta ecuación permite calcular la probabilidad de deserción estudiantil a partir de las variables finales incluidas en el modelo. Los coeficientes negativos de género y ocupación del estudiante señalan una disminución en la probabilidad estimada de deserción, mientras que los coeficientes positivos de estado civil 2 y estado civil "Otros" indican un incremento en dicha probabilidad.

La curva ROC permitió evaluar la capacidad del modelo para diferenciar entre estudiantes desertores y no desertores. En la Figura 3 se observa la curva obtenida para el modelo de regresión logística.

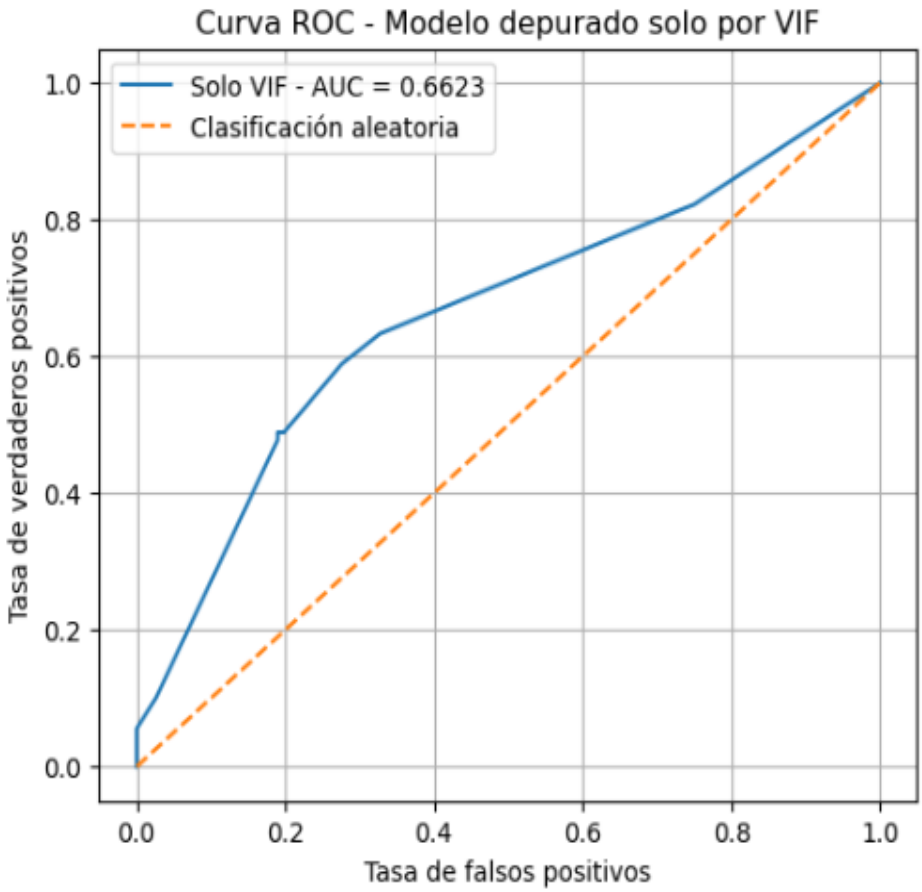

Figura 3. Curva ROC del modelo de regresión logística.
Elaboración: Los autores.

El área bajo la curva fue de AUC = 0,6623, lo que indica un desempeño moderado. Este valor muestra que el modelo logra clasificar mejor que una predicción al azar; sin embargo, todavía presenta limitaciones para identificar con alta precisión a todos los estudiantes en riesgo de deserción.

La prueba de Wald permitió evaluar la significancia de las variables incluidas en el modelo de regresión logística. En la Tabla 8 se presentan los coeficientes, valores de significancia y Odds Ratio obtenidos.

Tabla 8.
Prueba de Wald y Odds Ratio del modelo de regresión logística.

Variable	Coeficiente (B)	Error estándar	Wald	p-valor	Odds Ratio	IC 95% Inferior OR	IC 95% Superior OR
Constante	0.4737	0.1468	10.4163	0.0012	1.6060	1.2045	2.1413
Género_2	-1.0434	0.1754	35.3853	0.0000	0.3523	0.2498	0.4968
Ocupación del estudiante_1	-0.3510	0.1713	4.2002	0.0404	0.7040	0.5033	0.9848
Estado civil agrupado_Estado civil 2	0.4307	0.2575	2.7979	0.0944	1.5384	0.9287	2.5484
Estado civil agrupado_Otros	1.0817	0.3742	8.3584	0.0038	2.9498	1.4168	6.1417

Elaboración: Los autores.

Según los resultados, **Género_2**, **Ocupación del estudiante_1** y **Estado civil agrupado_Otros** presentaron significancia estadística dentro del modelo. En cambio, **Estado civil agrupado_Estado civil 2** obtuvo un p-valor de 0,0944, por lo que no fue significativo al nivel de 0,05, aunque mostró una tendencia positiva que debe interpretarse con cautela.

El género mostró un efecto protector, ya que se asoció con una menor probabilidad estimada de deserción. De igual manera, la ocupación del estudiante presentó una

relación significativa con el estado del estudiante, lo que indica que esta condición puede influir en la permanencia académica.

Por otra parte, la categoría "Otros" del estado civil agrupado reflejó un aumento en la probabilidad estimada de deserción en comparación con la categoría base. Esto sugiere que ciertos grupos dentro de esta variable podrían presentar mayor vulnerabilidad frente al abandono académico. La categoría "Estado civil 2" también presentó un coeficiente positivo, pero al no alcanzar significancia estadística al 5 %, su lectura debe considerarse solo como una posible tendencia.

Los Odds Ratio complementaron la interpretación del modelo, al mostrar qué variables disminuyen o incrementan la razón de probabilidades de deserción.

En conjunto, los resultados señalan que el modelo de regresión logística presentó un desempeño moderado en la clasificación del estado del estudiante. Aunque el análisis bivariado mostró una relación importante de las variables académicas con la deserción, el proceso de depuración por multicolinealidad facilitó ajustar el modelo final con un conjunto más reducido de variables. De esta manera, el género, la ocupación del estudiante y principalmente la categoría "Otros" del estado civil agrupado aportaron información relevante dentro del modelo. Estos hallazgos permiten considerar la regresión logística como una herramienta inicial de apoyo para el seguimiento institucional, aunque sus resultados deben interpretarse con cautela debido a las limitaciones observadas en la identificación de todos los casos de deserción.

DISCUSIÓN

Los resultados obtenidos señalan que la deserción estudiantil está asociada con varios factores y no responde a una sola causa. En el análisis bivariado se observó que las variables académicas, como el promedio, la repetición de materias, las asignaturas perdidas y el porcentaje de asistencia, presentaron una relación importante con el estado

del estudiante. Esto confirma que el rendimiento académico constituye un elemento relevante para comprender la permanencia o el abandono en la educación superior.

No obstante, al aplicar el análisis de multicolinealidad mediante VIF, varias de estas variables académicas presentaron valores elevados, lo que evidenció que compartían información similar dentro del modelo. Por esta razón, no todas fueron conservadas en la regresión logística final. Este resultado no significa que las variables académicas carezcan de importancia, sino que, al ser evaluadas junto con otros predictores, algunas aportaban información redundante y podían afectar la estabilidad de los coeficientes.

El modelo final quedó compuesto por género, ocupación del estudiante y estado civil agrupado. Sin embargo, dentro del estado civil agrupado, la categoría "Estado civil 2" no alcanzó significancia estadística al nivel de 0,05, aunque presentó una tendencia positiva. En cambio, la categoría "Otros" sí mostró significancia estadística y un incremento en la probabilidad estimada de deserción. Esto indica que ciertas condiciones personales o familiares pueden relacionarse con una mayor vulnerabilidad frente al abandono académico.

La permanencia del género y la ocupación del estudiante en el modelo demuestra que las condiciones personales y socioeconómicas también aportan información relevante para explicar la deserción estudiantil. En este sentido, la ocupación del estudiante puede vincularse con responsabilidades externas a la formación académica, las cuales podrían afectar el tiempo disponible para el estudio y la continuidad en el proceso formativo.

La estimación matemática del modelo permitió comprobar que los coeficientes obtenidos mediante Python responden al mismo fundamento de la regresión logística binaria, basado en la función logit y en el ajuste iterativo de los coeficientes. Esto refuerza la validez del modelo, ya que los resultados no dependen únicamente de la ejecución computacional, sino de un procedimiento estadístico sustentado en la estimación de probabilidades y en la evaluación del aporte de cada variable.

El desempeño del modelo fue moderado, con una exactitud de 0,665 y un AUC de 0,6623. Estos valores señalan que la regresión logística consiguió diferenciar entre estudiantes desertores y no desertores mejor que una clasificación al azar, aunque todavía presenta limitaciones. Además, el recall de la clase desertor fue de 0,49, lo que significa que el modelo identificó aproximadamente a la mitad de los estudiantes que realmente desertaron. Por ello, sus resultados deben tomarse como una herramienta inicial de apoyo y no como un sistema definitivo de predicción.

En general, los hallazgos permiten señalar que la regresión logística es útil para analizar la deserción estudiantil, especialmente porque permite interpretar el efecto de cada variable dentro del modelo. Sin embargo, para mejorar su capacidad predictiva en futuros estudios, sería conveniente incorporar nuevas variables relacionadas con seguimiento académico por período, historial de matrícula, carga laboral, apoyo familiar y uso de servicios institucionales. Esto permitiría fortalecer el modelo y ofrecer una mejor base para la toma de decisiones orientadas a la permanencia estudiantil.

CONCLUSIONES

El análisis bivariado permitió detectar que la deserción estudiantil se relaciona con variables académicas, sociodemográficas y socioeconómicas. Entre las variables académicas sobresalieron el promedio, la repetición de materias, las asignaturas perdidas y el porcentaje de asistencia, lo que evidencia que el rendimiento académico tiene un papel importante en el estudio de la permanencia estudiantil.

El análisis de multicolinealidad mediante VIF permitió depurar el conjunto inicial de variables y reducir la redundancia entre predictores. Aunque varias variables académicas fueron significativas en el análisis inicial, no todas se conservaron en el modelo final debido a que compartían información similar. De esta manera, el modelo final quedó conformado por género, ocupación del estudiante y estado civil agrupado.

La estimación matemática de la regresión logística permitió sustentar los coeficientes obtenidos en el modelo final. Este procedimiento confirmó que los resultados generados mediante Python corresponden al fundamento estadístico de la regresión logística binaria, basado en la función logit y en la estimación iterativa de los coeficientes.

El modelo presentó un desempeño moderado, con una exactitud de 0,665 y un AUC de 0,6623. Estos resultados señalan que el modelo consigue diferenciar entre estudiantes desertores y no desertores mejor que una clasificación al azar; sin embargo, todavía presenta limitaciones para reconocer todos los casos de deserción.

La prueba de Wald y los Odds Ratio permitieron interpretar el aporte de las variables finales dentro del modelo. Género_2, Ocupación del estudiante_1 y Estado civil agrupado_Otros presentaron significancia estadística, mientras que Estado civil agrupado_Estado civil 2 no fue significativo al nivel de 0,05, aunque mostró una tendencia positiva. En términos generales, el género y la ocupación del estudiante se asociaron con una disminución en la probabilidad estimada de deserción, mientras que la categoría "Otros" del estado civil agrupado se relacionó con un incremento de dicha probabilidad.

Finalmente, se concluye que la regresión logística puede emplearse como una herramienta inicial de apoyo para el seguimiento institucional de la deserción estudiantil. No obstante, sus resultados deben complementarse con criterios académicos y administrativos antes de tomar decisiones, especialmente porque el modelo no identifica la totalidad de estudiantes en riesgo.

FINANCIAMIENTO

No monetario.

AGRADECIMIENTO

A los directivos y colaboradores del Instituto Superior Tecnológico General Eloy Alfaro por su valioso apoyo en el desarrollo de esta investigación.

REFERENCIAS CONSULTADAS

Abdulkareem Shafiq, D., Marjani, M., Ariyaluran Habeeb, R., & Asirvatham, D. (2022). Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review. *10*. https://n9.cl/y0zl6k

Aguilar Reyes, J. E., Mejía Peñafiel, E. F., Morocho Barrionuevo, T. P., & Velasco Castelo, G. M. (2025). Estudio del rendimiento académico mediante la comparación de modelos de regresión y árboles de clasificación. *Telos*, *27*(1), 94-115. https://n9.cl/l6vpcu

Alcaraz González, R., Morales Benítez, B., González García, M., & Paredes Medina, I. (23 de 9 de 2024). Modeling School Dropout at the Faculty of Tourism of UAGro through Logistic Regression. *Revista RELEP – Educación y Pedagogía en Latinoamérica, 6*(3), 17–30. https://n9.cl/d7bjb

Aulck, L., Velagapudi, N., Blumenstock, J. E., & West, J. (2016). Predicting Student Dropout in Higher Education. https://n9.cl/fg8t7

Cárdenas Matute, J. M., Valle Franco, A., & Tapia Segarra, J. I. (2023). Factores que inciden en la deserción estudiantil en la unidad académica de Ciencias Sociales de la Universidad Católica de Cuenca. *ConcienciaDigital, 6*(3), 30-48. https://n9.cl/kk7ewm

Carvajal, C. M., González, J. A., & Sarzoza, S. J. (2018). Variables Sociodemográficas y Académicas Explicativas de la Deserción de Estudiantes en la Facultad de Ciencias Naturales de la Universidad de Playa Ancha (Chile). *Formación universitaria*, *11*(2), 3-12. https://n9.cl/likymg

Fernández Casal, R., Costa Bouzas, J., & Oviedo de la Fuente, M. (2024). Métodos predictivos de aprendizaje estadístico. https://n9.cl/jwrqt

Gutiérrez Monsalve, J. A., Garzón, J., & Segura Cardona, A. M.(2021). Factores asociados al rendimiento académico en estudiantes universitarios. *Formación universitaria*, *14*(1), 13-24. https://n9.cl/r4kkt

Hosmer Jr, D., Lemeshow, S., & Sturdivant, R. (2013). Applied Logistic Regression (3rd Edition). https://n9.cl/mem1hx

Morales, L. A., Xicoténcatl Ramírez, G., Ibarra Corona, M. A., & García Reyes, D. A. (2024). Factores y estrategias que influyen en la deserción en educación superior: Revisión sistemática. *RIDE Revista Iberoamericana para la Investigación y el Desarrollo Educativo, 15*(29). https://n9.cl/3cp3y

Nigri, A., Bilancia, M., & Cafarelli, B. (2025). Modelling higher education dropouts using sparse and interpretable post-clustering logistic regression. https://n9.cl/fi78x

Quincho Apumayta, R., Flores Poma, I., Salazar Mucha, W. C., Cárdenas Flores, K., Cárdenas Flores, J., & Goyas Baldoceda, A. M. (2025). Trayectorias truncadas: factores críticos en la deserción universitaria desde una perspectiva sistémica. *e-Revista Multidisciplinaria del Saber,3*. https://n9.cl/nlzm7

Rodríguez, A., Espinoza, J., Ramírez, L., & Ganga, A. (2018). Deserción Universitaria: Nuevo Análisis Metodológico. *Formación universitaria, 11(6),* 107-118. https://n9.cl/u901c

Salmerón Gómez, R., & Rodríguez Martínez, E. (2017). Métodos cuantitativos para un modelo de regresión lineal con multicolinealidad. Aplicación a rendimientos de letras del tesoro. *24*, 169-189. https://n9.cl/dr7r7

Salmerón, R., García, C., & García, J. (2020). Overcoming the inconsistencies of the variance inflation factor: A redefined VIF and a test to detect statistical troubling multicollinearity. https://n9.cl/r61ld

Villegas, R., Núñez, L., & Luis, A. (2024). Factores asociados a la deserción estudiantil en el ámbito universitario. Una revisión sistemática 2018-2023. *RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, *14*(28). https://n9.cl/ktxm7