PROMETEO
Conocimiento Científico
ISSN 2953-4275
Vol. 5, Núm. 1, enero-junio 2025, e96.

Educación
Artículo de Investigación Original

MODELO PREDICTIVO DE RENDIMIENTO ACADÉMICO BASADO EN TÉCNICAS MULTIVARIANTES

PREDICTIVE MODEL OF ACADEMIC PERFORMANCE BASED ON MULTIVARIATE TECHNIQUES

Blanca Rocio Cuji Chacha[1][0000-0003-4091-6876], Wilma Lorena Gavilánes López [2][0000-0002-2563-6633]

[1] Universidad Politécnica Estatal del Carchi, Posgrado. Antisana y Avenida Universitaria. 040101. Carchi-Tulcán. Ecuador
[2] Universidad Técnica de Ambato. Facultad de Ciencias Humanas y de la Educación. Avenida los Chasquis y Rio Payamino. 180103. Tungurahua - Ambato- Ecuador
blanca.cuji@upec.edu.ec wilmalgavilanesl@uta.edu.ec

CITA EN APA:

Cuji Chacha, B. R., & Gavilánes López, W. L. (2025). Modelo predictivo de rendimiento académico basado en técnicas multivariantes. *Prometeo Conocimiento Científico*, *5*(1). https://doi.org/10.55204/pcc.v5i1.e96

Recibido: 03 de enero 2025
Aceptado: 22 de mayo 2025
Publicado: 10 de junio 2025

Prometeo
Conocimiento Científico
ISSN: 2953-4275

Los contenidos de este artículo están bajo una licencia de Creative Commons Attribution 4.0 International (CC BY 4.0)

Los autores conservan los derechos morales y patrimoniales de sus obras.

Resumen.

El rendimiento académico está asociado a diversos factores de tipo sociales, pedagógicos, económicos, institucionales. El objetivo de esta investigación es diseñar un modelo predictivo de rendimiento académico de los estudiantes, para la asignatura de matemáticas. Para evaluar el impacto de este trabajo se aplicaron técnicas multivariantes como la regresión lineal múltiple. Se concluye que las variables, Componente Aprendizaje Autónomo, Aprendizaje Practico Experimental, Aprendizaje en Contacto con el Docente, y la edad, inciden significativamente en el rendimiento académico. Es necesario incluir más variables para tener un modelo robusto que permita, mejorar el desempeño y permanencia de los estudiantes en una institución de educación superior.

Palabras Clave: Rendimiento académico, técnicas multivariantes, regresión lineal multiple.

Abstract:

Academic performance is associated with various social, pedagogical, economic, and institutional factors. The objective of this research is to design a predictive model of students' academic performance for the subject of mathematics. To evaluate the impact of this work, multivariate techniques such as multiple linear regression are applied. It is concluded that the variables, Autonomous Learning Component, Experimental Practical Learning, Learning in Contact with the Teacher, and age, significantly affect academic performance. It is necessary to include more variables to have a robust model that allows improving the performance and permanence of students in a higher education institution.

Keywords: Academic performance, multivariate techniques, multiple linear regression.

1. INTRODUCIÓN

El propósito del rendimiento académico es alcanzar un aprendizaje. Por tal razón, sus componentes son variados y pueden estar determinados, de acuerdo con las circunstancias y condiciones ambientales que determinan las aptitudes y experiencias de los estudiantes. El rendimiento académico puede convertirse en una problemática, que afecta, tanto a estudiantes, profesores y autoridades, de una institución educativa (Lamas, 2014).

En la Facultad de Contabilidad y Auditoría de la Universidad Técnica de Ambato (UTA), se ha identificado un problema general, relacionado con el bajo rendimiento de los estudiantes de la carrera de Contabilidad, en la asignatura de Estadística. Siendo el rendimiento académico una de las causas más comunes de la deserción estudiantil (Ortega et al., 2022), es necesario su análisis individual. Se dispone de escasa información, y las existentes, mayormente están referidas a investigaciones de aspectos específicos como: calificaciones por promoción, motivaciones para estudiar la carrera, perfil profesional. No se evidencia a nivel institucional y de facultad un mecanismo, debidamente organizado y normado, para el análisis predictivo del bajo rendimiento académico en asignaturas como la estadística o matemática. Por esta razón, se presenta la situación problémica expuesta anteriormente.

Existen diversas técnicas multivariantes para la predicción del rendimiento académico, entre ellas tenemos los modelos lineales, regresión lineal múltiple. Además, técnicas como arboles de decisión, regresión lineal cada una de ellas con sus propias características.

Este estudio se centra en los modelos predictivos de rendimiento académico, utilizando técnicas enfocadas en modelos lineales, por ajustarse a los datos y como parte de un trabajo futuro propuesto en otras investigaciones previas.

El estudio relacionado con “Modelo Predictivo del rendimiento académico de estudiantes de Ingeniería Química en el área de Matemáticas”, presenta un modelo predictivo del desempeño de los estudiantes de Ingeniería Química del Instituto Tecnológico de Pachuca en el área de matemáticas, tomando como datos los resultados del examen de ingreso. La investigación es cuantitativa, con enfoque correlacional y descriptivo. Se toman los datos de 380 estudiantes a partir del año 2012 al 2016. Se determina la correspondencia directa entre el promedio obtenido en las materias y los puntos en las áreas de matemáticas del examen, se realiza una análisis comparativo de siete materias del área de matemáticas, logrando diseñar un modelo predictivo con un grado de confiabilidad aceptable (Castelazo et al., 2022).

Por otra parte, (Herreras, 2019), se centra en analizar y predecir el rendimiento matemático en función de atribuciones al fracaso y el enfoque de aprendizaje, mediante la aplicación de un cuestionario se levantan de 25313 jóvenes de 15 años, utilizando la prueba de muestreo por conglomerados. Se toma en cuenta aspectos como los procesos, contenidos matemáticos, metodología aplicada por el profesor, interés de los estudiantes por la asignatura, entre otros. Se utiliza regresión logística binaria para estimar el riesgo de tener un rendimiento (bajo, medio y alto) en función de: atribuciones al fracaso y enfoques de aprendizaje

(Herreras, 2019). Además, se utiliza la prueba de Hosmer y Lemeshow para la comprobación de hipótesis y las pruebas de R2 (Cox y Snell, Nagelkerke) para determinar que no se puede predecir el bajo rendimiento en matemáticas a partir de la atribución causal y el enfoque de aprendizaje dos variables que forman parte del estudio.

Para (Dorta Guerra et al., 2019), en su estudio relacionado con "Un modelo predictivo del rendimiento académico a partir de las calificaciones de Bachillerato y PAU", se menciona al rendimiento académico como una forma de medir la calidad de la educación, por tanto se considera importante determinar variables que afectan al rendimiento. Se diseña un modelo predictivo utilizando regresión lineal múltiple, obtenido como resultado que la variable predictora dominante es la nota media de Bachillerato y que el modelo encontrado se convierte en una herramienta que puede ser empleada por la universidad para detectar aquellos individuos susceptibles de ser orientados hacia la mejora de su aprovechamiento académico.

Bajo este contexto el objetivo de la investigación se centró en diseñar un modelo predictivo de rendimiento académico, para la asignatura de estadística de los estudiantes de la Carrera de Contabilidad y Auditoría de la UTA, a través de realizar un diagnóstico de los factores que afectan al rendimiento académico, posteriormente se elaboró un modelo predictivo de rendimiento académico, basado en técnicas multivariantes, y se evaluó para medir su efectividad.

Por medio del cumplimiento del objetivo se dio respuesta a las preguntas de la investigación planteadas como: ¿Cuáles son los factores que afectan al rendimiento académico de los estudiantes de la carrera de Contabilidad y Auditoría en la asignatura de estadística?, ¿En qué medida la creación de un modelo predictivo basado en técnicas multivariantes facilitará la toma de decisiones en relación al rendimiento académico? y ¿Cuáles son los resultados de las pruebas realizadas al modelo predictivo de rendimiento académico?. La hipótesis de la investigación, se centró en comprobar si el diseño de un modelo predictivo de rendimiento académico basado en diferentes factores, permite precautelar el bajo rendimiento académico.

Justificación

Considerando que el rendimiento académico, es una característica esencial en las instituciones educativas, por ser uno de los principales criterios para medir la calidad académica (Cabrera et al., n.d.). Alcanzar a predecirlo, representa un beneficio, para profesores y estudiantes, lo que implica, proponer programas de prevención para estudiantes con bajo rendimiento y localizar previamente aquellos en peligro de deserción.

La inversión en la gratuidad de la educación puede representar un perjuicio para el estado. Carlos Cedeño, rector de la Universidad de Guayaquil manifiesta: "Apenas el 30% de los bachilleres que ingresan a la universidad logran graduarse, en tanto, que un 70% se va quedando en el camino". De aquí, parte la importancia del estudio del rendimiento académico, como uno, de los factores que inciden en la deserción,

y afectan directamente a la gratuidad de la educación, lograr disminuir los índices de deserción en las instituciones educativas, genera un impacto económico, en el presupuesto anual del estado.

Por otra parte, la investigación aporto al cumplimiento del objetivo Estratégico Institucional # 1 de la UTA que manifiesta: "Formar y especializar profesionales con liderazgo, responsabilidad social ambiental con sólidos conocimientos científicos, tecnológicos y artísticos, que entiendan la realidad socioeconómica del Ecuador".

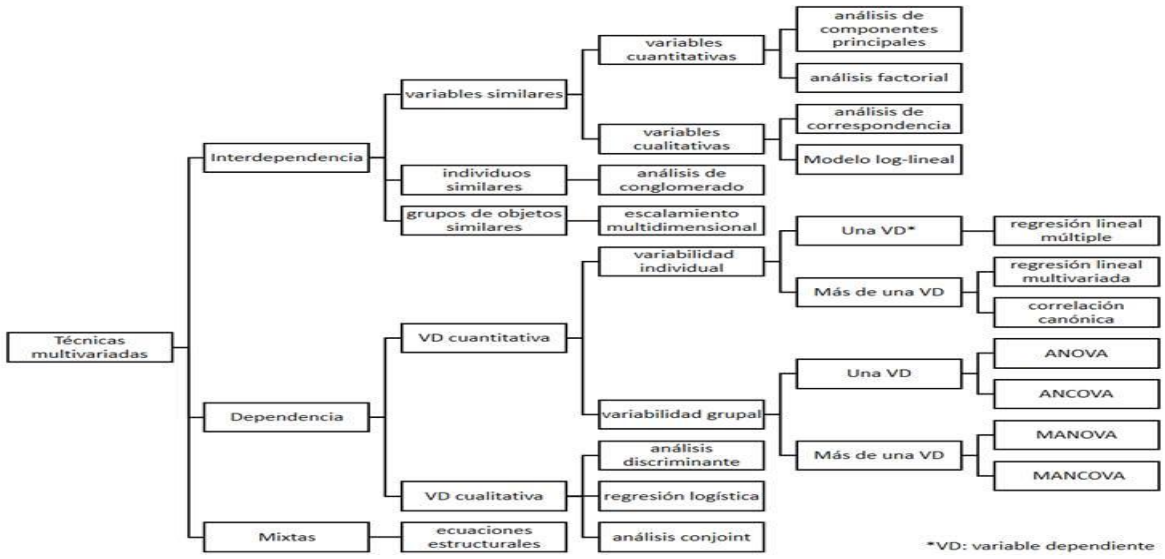
Para llevar a cabo la investigación, se contó con recursos tecnológicos, como un computador de última generación y software estadístico libre R disponible bajo una Licencia Pública, además de PSPP, que permitió poner en práctica diferentes métodos estadísticos, como modelos lineales basados en técnicas de regresión.

Los beneficiarios directos fueron 90 estudiantes del tercer semestre de la Carrera de Contabilidad y Auditoría de la UTA, así como, cuatro docentes del área de ciencias exactas. Se menciona, además, como beneficiarios indirectos aproximadamente 180 estudiantes de Carrera, docentes y autoridades de Facultad.

Técnicas Multivariantes

Son aquellas que analizan múltiples características medidas en un mismo individuo, que por estar interrelacionadas no tiene sentido medir su efecto de manera aislada.

Figura 1: Clasificación de las Técnicas Multivariantes

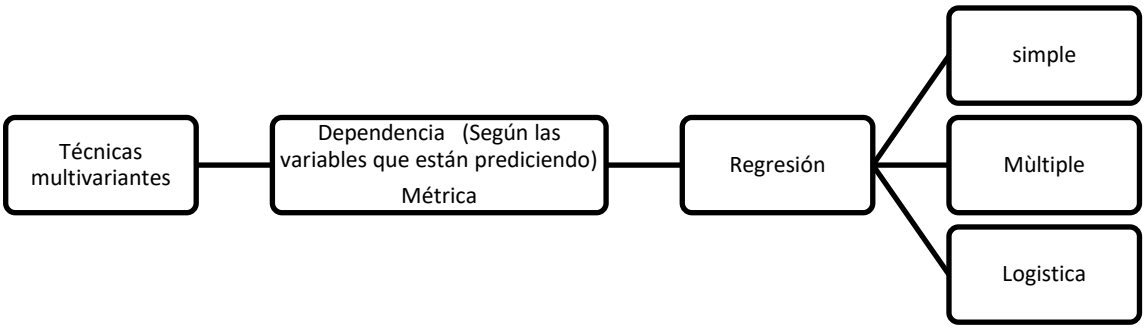


Fuente: (Campo & Matamoros, 2020)

Regresión múltiple, es un tipo de técnica multivariada (Hair et al., 2019), que permite generar un modelo lineal en función de una variable dependiente o respuesta (y), que se determina a partir de un conjunto de variables independientes llamadas predictores (X1,X2,X3,.., Xn). Es una extensión de la regresión lineal simple (que consta de una variable explicada y una variable explicativa), estos modelos pueden emplearse para predecir el valor de la variable dependiente en función de variables independientes.

En este tipo de modelos es importante testar la heterocedasticidad, la multicolinealidad, especificación (Granados, 2016). La Figura 2 muestra una segregación de la técnica multivariante a utilizar.

Figura 2: Regresión Lineal Múltiple



Fuente: Investigador

Modelo de regresión lineal múltiple (MRL)

Un modelo de regresión lineal múltiple expresa una relación estadística, entre una variable "x" y "y", describe la variación de "y" con respecto a "x". Dónde: y: Es una variable de respuesta o dependiente (continua) y, $x_1, \dots, x_k$ son k, variables predictores, auxiliares, regresoras o independientes.

La regresión se aplica con frecuencia a datos observacionales (las variables X no está controladas). Para analizar la regresión lineal múltiple se debe considerar los siguientes aspectos:

1. Validez y ajuste del modelo.
2. Ecuación de regresión.
3. Análisis de los supuestos.

La ecuación de la regresión lineal con múltiples variables es:

$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n + e$

Y: variable dependiente

b_0: es la ordenada en el origen, el valor de la variable dependiente.

b_1x_1: representa el coeficiente de regresión (b_1) sobre la primera variable independiente (x_1). El mismo análisis se aplica a todos los coeficientes y variables de regresión restantes.

e es el error del modelo (residuales), que define cuánta variación se introduce en el modelo al estimar Y.

Los supuestos de regresión lineal múltiple.

1. **Normalidad**: Se comprueba utilizando un gráfico de probabilidad normal o un histograma.
2. **Linealidad:** Debe haber una relación lineal entre las variables dependientes e independientes. Se puede ilustrar mediante diagramas de dispersión que muestran una relación lineal o curvilínea.
3. **Multicolinealidad**: las variables independientes no están altamente correlacionadas entre sí. La multicolinealidad dificulta identificar qué variables explican mejor la variable dependiente. Esta suposición se verifica calculando una matriz de correlaciones bivariadas de Pearson entre todas las variables independientes. Si no hay colinealidad en los datos, todos los valores deben ser inferiores a 0,8.
4. **Homocedasticidad** asume que la varianza de los errores residuales es similar a través del valor de cada variable independiente.

https://doi.org/10.55204/pcc.v5i1.e96

En general la finalidad de la regresión múltiple, es obtener los valores de los parámetros de regresión, de tal forma que la suma de los cuadrados de los errores o residuos sea mínima, con el propósito de optimizar la predicción(Tortolero Portugal et al., 2020)

Rendimiento académico

El rendimiento académico se relaciona con el nivel de logro o éxito, que un estudiante puede alcanzar en una o varias asignaturas, es una característica esencial en las instituciones educativas, por ser uno de los principales criterios para medir la calidad académica (Rico Páez et al., 2019). Alcanzar a predecir el rendimiento académico, representa un beneficio, para profesores y estudiantes, lo que implica, proponer programas de prevención para estudiantes con bajo rendimiento, localizar previamente a estudiantes en peligro de deserción, etc. Ademas, permite evaluar los resultados de un curso sobre el conjunto de asignaturas de una entidad educativa, muchas veces se mide exclusivamente, por medio de las calificaciones lo que implica un riesgo debido, fundamentalmente a la subjetividad de los docentes, sin embargo las calificaciones no dejan de ser el medio más usado para cuantificar el rendimiento académico (Fernández, 2021).

El rendimiento académico depende de varios factores que pueden o no estar relacionados. De acuerdo a (Collazos et al., 2021) el rendimiento se basa en cinco categorías: personales, familiares, académicas, económicas e institucionales, las cuales se miden en un enfoque cuantitativo y cualitativo. La valoración del rendimiento académico permite identificar la relación entre lo que el estudiante aprende y lo que alcanza en el proceso de enseñanza – aprendizaje como producto del trabajo académico en las diferentes actividades en las que se desempeñó ante sus maestros (Naal et al., 2022).

Factores que influyen en el rendimiento académico

- Factores Psicológicos

Estos factores muestran el papel que se le atribuye a la depresión en la limitación del rendimiento académico, así como la ansiedad ante los exámenes, disminuye la capacidad del desempeño académico. Además, se menciona a los síntomas depresivos, como una de las causas del bajo rendimiento de los estudiantes (Vázquez et al., 2022).

- Factores Sociales

Estos factores se relacionan, con el estrato social, situación intrafamiliar que afectan directamente a la concentración, atención y motivación de los/as estudiantes, por estudiar, estrategias que utilizan los padre para estimular a sus hijos a estudiar interacción entre padres e hijos, tipo de educación del nivel medio(pública, privada) (Collazos et al., 2021).

- Factores Académicos

Las variables académicas que se asocian al rendimiento académico, abarcan aquellos factores educativos relacionados con la intervención pedagógica o docente. En consecuencia, las variables pedagógicas incorporan lo interno: el esfuerzo del profesor; en tanto que las variables académicas tienden

a incluir lo externo a la práctica docente: el esfuerzo del alumno y la política académica de la institución (Fernández, 2021).

- Factores personales

Las variables personales generalmente relacionadas, con el rendimiento académico están: las habilidades de estudio, la organización y concentración, la capacidad para relacionar nuevos conocimientos con los existentes, la comprensión lectora y la capacidad para autorregular el aprendizaje (Amín & Amezcua, 2021).

- Factores económicos

Los factores económicos asociados al rendimiento académico vienen dados por el nivel de escolaridad, calidad de vida y oportunidades, ingresos de los padres (Collazos et al., 2021).

2. METODOLOGÍA O MATERIALES Y METODOS

Enfoque

La investigación tuvo un enfoque cuantitativo para describir variables relacionadas con el rendimiento académico como el promedio general, Componente Aprendizaje Autónomo, Aprendizaje Practico Experimental, Aprendizaje en Contacto con el Docente, y la edad, edad. Además, se toma en cuenta variables como: etnia, estado civil, ciudad de nacimiento, residencia, entre otras.

Para el análisis y descripción de las variables tanto cuantitativas como cualitativas se utilizó métodos científicos como: el método histórico-lógico, para analizar la trayectoria del rendimiento académico de los estudiantes en la asignatura de estadística desde el inicio del semestre hasta su finalización, función del análisis histórico describir de forma ordenada y lógica las causas del bajo rendimiento de los estudiantes.

Puesto que la investigación tiene un enfoque cuantitativo fue necesario la recolección de datos numéricos, para obtener resultados, que permitan probar la hipótesis plateada, este proceso requiere objetividad, a través del uso de un razonamiento deductivo, permitiendo ir de un análisis general del rendimiento académico, algo particular como la determinación de los factores que inciden en el mismo. Por otra parte, el enfoque cualitativo permite la recolección de datos sin medición numérica, y la elaboración de preguntas de investigación, por medio de un razonamiento inductivo, en función de las variables seleccionadas en forma particular, llevarán a la creación de un modelo predictivo general de rendimiento académico y por este hecho será de tipo explicativa y transversal.

Además, para el análisis de los datos, se aplicó métodos estadísticos, como modelos lineales, enfocados en regresión múltiple, para obtener un modelo predictivo del rendimiento académico en función las variables seleccionadas.

Unidades de análisis

Los datos pertenecen a los estudiantes de tercer nivel de la Carrera de Contabilidad y Auditoría de la UTA, contiene datos sobre: Componente Aprendizaje Autónomo, Aprendizaje Practico Experimental,

Aprendizaje en Contacto con el Docente, edad. Además, datos de tipo etnográficos como: género, estado civil, etnia, fecha de nacimiento, ciudad de nacimiento y de residencia.

Técnicas e instrumentos de investigación

Las técnicas que se utilizaron durante la investigación fueron, la observación y el análisis de la base de datos histórica, permitiendo la identificación de los factores que causan un bajo rendimiento en la asignatura de estadística, durante el período 2022-2023.

Procesamiento de análisis

Posterior a la determinación de los factores, que afectan directamente al rendimiento académico de los estudiantes, en la asignatura de estadística, se utilizó técnicas multivariantes para el diseño del modelo predictivo. La técnica a utilizar se enfoca en los modelos lineales como la regresión múltiple. Con ayuda del Software RStudio se obtuvo un modelo predictivo de rendimiento académico.

La base de datos, originalmente contenían datos faltantes NA's, y outliers, previo a la aplicación de la técnica multivariante de regresión lineal múltiple, fue necesario realizar una fase de limpieza de datos para obtener información significativa de los datos. Inicialmente se utilizaron *bloxplot* para cada variable donde se detectaron los aoutliers, se procedió a eliminarlos puesto que solo representaban el 2% de los datos. Para los N/A se reemplazaron con el valor de la media aritmética de los datos originales.

La metodología de trabajo, se divide en cuatro fases partiendo del análisis de la información histórica de una base de datos de 270 estudiantes de la carrera de Contabilidad y Auditoría 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 3).

Figura 3: Etapas de la metodología.

a. Análisis variables

En esta fase se realizó la recolección e integración de datos, que pertenece a 285 estudiantes. La institución proporciono los datos crudos a partir del año 2022 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: Componente Aprendizaje Autónomo, Aprendizaje Practico Experimental, Aprendizaje en Contacto con el Docente, y la edad.

Además, datos de tipo etnográficos como: género, estado civil, etnia, fecha de nacimiento, ciudad de nacimiento y de residencia.

Las variables con las que se cuenta para la generación del modelo se clasificaron (Tabla I) (Clifford & Taylor, 2014) en:

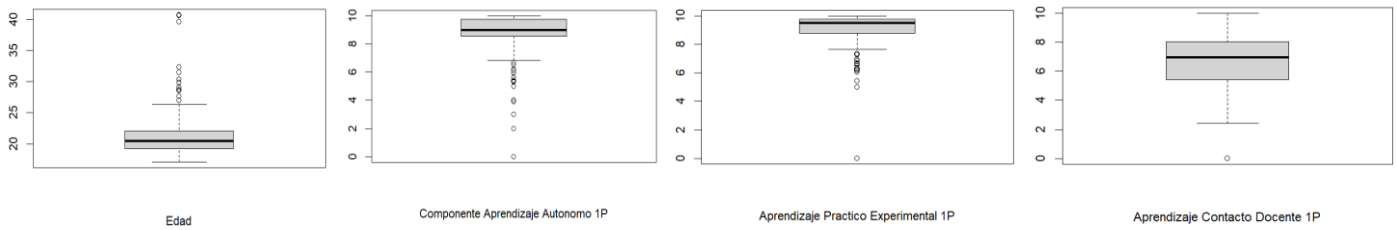
- Nominal (los valores son básicamente una función de etiquetado).
- Ordinal (los valores obedecen a la relación de orden).
- Cuantitativo (los valores tienen todo el poder expresivo de los números reales).

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
Edad (ED)	Cuantitativo	
Lugar de Nacimiento (LN)	Cualitativa – Nominal dicotómica	Ambato Otros
Ciudad de residencia (CR)	Cualitativa – Nominal dicotómica	Ambato Otros
AA_Componente_Aprendizaje_Autonomo_1P (AA_1P)	Cuantitativo	
APE_Aprendizaje_Practico_Experimental_1P (APE_1P)	Cuantitativo	
ACD_Aprendizaje_Contacto_Docente_1P (ACD_1P)	Cuantitativo	
AA_Componente_Aprendizaje_Autonomo_2P (AA_2P)	Cuantitativo	
APE_Aprendizaje_Practico_Experimental_2P (APE_2P)	Cuantitativo	
ACD_Aprendizaje_Contacto_Docente_2P (ACD_2P)	Cuantitativo	
Promedio_General	Cuantitativo	
Asistencia_General	Cuantitativo	

b. Limpieza de Datos

En primera instancia se detectan datos anómalos a través de un diagrama un bloxplot. Para las variables predictores (EDAD, AA1P, APE1P, ACD1P, AA2P, APE2P, ACD2P).



Identificación de outsiders para las variables:

Edad: 32,31,29,39,30,40,29,28,26,40,27,29,28

AA1P: 5.0,3.0,5.0, 6.5, 6.5, 0.0, 6.5, 6.5, 3.9, 6.1, 5.6, 5.9,5.4, 6.1, 0.0, 4.0, 0.0, 0.0, 4.0, 6.5, 2.0, 4.0, 4.0, 6.7, 5.3, 6.2, 5.4

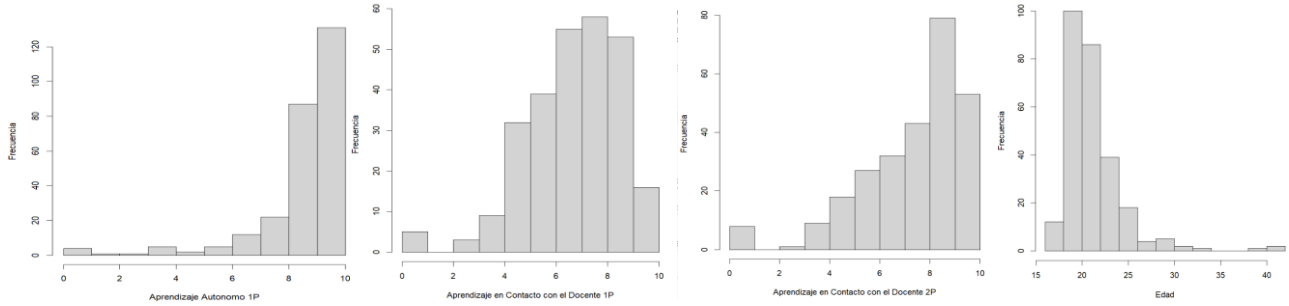
APE1P: 0.0, 6.7, 6.7, 6.5, 6.2, 6.5, 0.0, 6.2, 6.3, 0.0, 0.0, 5.4, 5.4, 6.9, 0.0, 6.1, 5.0, 0.0, 7.2, 6.6, 6.7, 6.9, 6.5, 6.3, 7.3, 7.3

ACD1P: 0,0,0,0,0

Se procede de la misma forma para las variables AA2P, APE2P, ACD2P. El 2% de los datos detectados como outsiders, fueron eliminados de la base de datos.

c. Formulación del modelo predictivo de regresión lineal múltiple

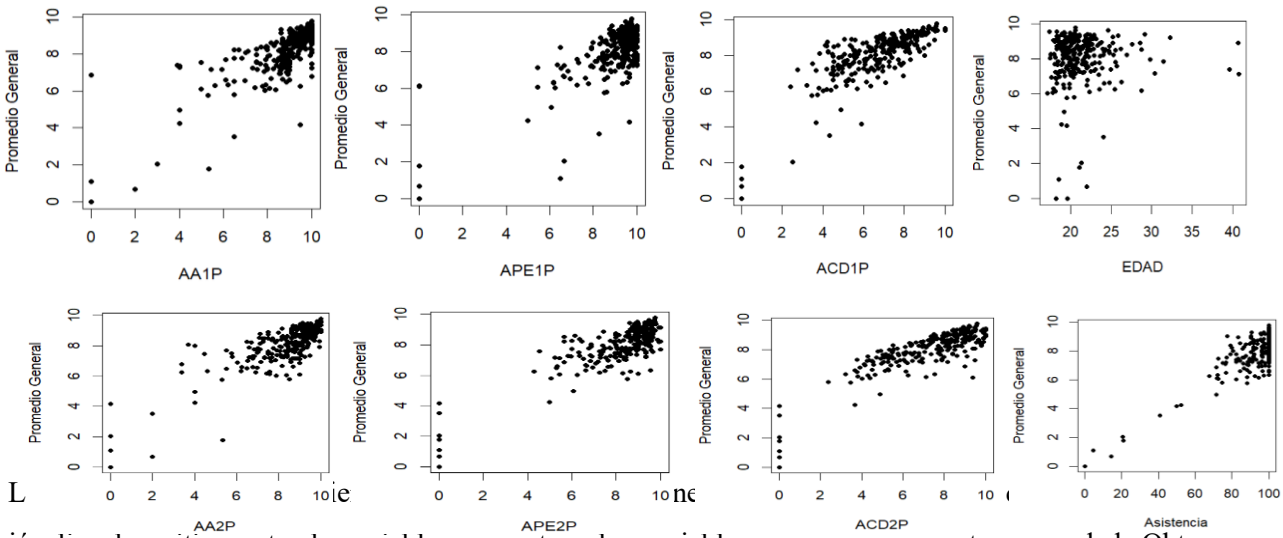
1. Inicialmente se comprueba la normalidad de los datos utilizando un histograma.



Se analiza la normalidad de las variables regresoras AA1P, ACD1P, ACP2P, EDAD.

2. Se analiza la linealidad de los datos, mediante diagramas de dispersión. Existe una elevada dispersión en la variable edad, y aprendizaje en contacto con el docente del segundo parcial, en relación al valor medio (Ver gráfico 1).

Gráfico 1: Diagrama de dispersión de las variables regresoras en relación a la variable respuesta



Linealidad

correlación lineal positiva entra la variable respuesta y las variables regresoras, excepto para edad. Obtenemos entonces la correlación lineal de Pearson, presentando únicamente la correlación existente entre las variables regresoras y la variable respuesta (Promedio_General), (Ver Tabla 3).

Tabla 2: Correlación lineal de Pearson de la variable respuesta con las variables regresoras

	PROMEDIO_GENERAL	EDAD	AA1P	APE1P	ACD1P	AA2P	APE2P	ACD2P	ASISTENCIA_GENER
PROMEDIO_GENERAL	1.00000000	0.033599507	0.779204392	0.73439364	0.78151237	0.81302539	0.84121558	0.84044541	0.8586027

Al igual que en la forma gráfica se demuestra que existe alta correlación entre la variable respuesta y las regresoras, excepto con la variable edad cuya correlación es baja.

3. **Multicolinealidad**: Esta suposición se verifica calculando una matriz de correlaciones bivariadas de Pearson entre todas las variables independientes. Si no hay colinealidad en los datos, todos los valores deben ser inferiores a 0,8 (Ver tabla 4).

Tabla 3 Correlación de las variables regresoras o independientes:

	EDAD	AA1P	APE1P	ACD1P	AA2P	APE2P	ACD2P	ASISTENCIA_GENER
EDAD	1.000000000	0.003788162	0.02365716	0.04368526	0.07447594	-0.02373031	0.03715375	0.0395629
AA1P	0.003788162	1.000000000	0.56540213	0.53840479	0.57730747	0.58940438	0.51107351	0.6972665
APE1P	0.023657161	0.565402127	1.000000000	0.51690628	0.44454204	0.56735894	0.49053830	0.6555812
ACD1P	0.043685265	0.538404789	0.51690628	1.000000000	0.51357605	0.50053109	0.65876750	0.5805778
AA2P	0.074475944	0.577307466	0.44454204	0.51357605	1.000000000	0.71641027	0.63774138	0.7538586
APE2P	-0.023730314	0.589404375	0.56735894	0.50053109	0.71641027	1.000000000	0.65445277	0.7954991
ACD2P	0.037153749	0.511073512	0.49053830	0.65876750	0.63774138	0.65445277	1.000000000	0.6479376
ASISTENCIA_GENERAL	0.039562901	0.697266549	0.65558120	0.58057782	0.75385862	0.79549913	0.64793759	1.0000000

d. Creación del modelo

Se ajusta el modelo de regresión lineal múltiple con todas las variables regresoras expuestas, obteniendo el siguiente modelo predictivo inicial de rendimiento académico

lm(formula = PROMEDIO_GENERAL ~ AA1P + APE1P + ACD1P + AA2P + APE2P + ACD2P + EDAD + ASISTENCIA_GENERAL, data = papercor)

Tabla 4: Modelo inicial de regresión lineal múltiple

(Intercept)	AA1P	APE1P	ACD1P	AA2P	APE2P	ACD2P	EDAD	ASISTENCIA_GENERAL
7.567e-15	1.667e-01	1.667e-01	1.667e-01	1.667e-01	1.667e-01	1.667e-01	1.988e-17	2.159e-17

Posterior al modelo inicial se aplica el método “paso a paso”, para seleccionar el mejor grupo de variables regresoras, mejorando el ajuste.

step(object = modelo, direction = "both", trace=1)

Donde:

object es el modelo obtenido originalmente, *direction*: va a sumar o restar las variables dependiendo de su importancia en el modelo, *trace:* muestra paso a paso la selección de variables para el modelo ajustado. El Gráfico 2

muestra que las variables “EDAD” y “ASISTENCIA_GENERAL”, han sido eliminadas. Con respecto al AIC (medida del rendimiento del modelo) que compara modelos de regresión, teniendo en cuenta la complejidad del modelo, el AIC más bajo proporciona un mejor ajuste de los datos observados.

Gráfico 2:Modelo ajustado

```
Start:  AIC=-18140.18
PROMEDIO_GENERAL ~ AA1P + APE1P + ACD1P + AA2P + APE2P + ACD2P +
    EDAD + ASISTENCIA_GENERAL

                     Df Sum of Sq     RSS      AIC
- EDAD                1    0.0000  0.0000 -18142.0
- ASISTENCIA_GENERAL  1    0.0000  0.0000 -18141.1
<none>                             0.0000 -18140.2
- APE2P               1    7.8938  7.8938   -937.7
- AA2P                1    8.6848  8.6848   -912.0
- APE1P               1   10.3298 10.3298   -865.1
- AA1P                1   10.3743 10.3743   -864.0
- ACD1P               1   12.5203 12.5203   -813.2
- ACD2P               1   13.9446 13.9446   -784.1
```

```
Step:  AIC=-18142
PROMEDIO_GENERAL ~ AA1P + APE1P + ACD1P + AA2P + APE2P + ACD2P +
    ASISTENCIA_GENERAL

                     Df Sum of Sq     RSS      AIC
- ASISTENCIA_GENERAL  1    0.0000  0.0000 -18142.9
<none>                             0.0000 -18142.0
+ EDAD                1    0.0000  0.0000 -18140.2
- APE2P               1    8.0038  8.0038   -936.0
- AA2P                1    8.7664  8.7664   -911.4
- APE1P               1   10.3349 10.3349   -867.0
- AA1P                1   10.3941 10.3941   -865.4
- ACD1P               1   12.5235 12.5235   -815.1
- ACD2P               1   13.9470 13.9470   -786.1
```

```
Step:  AIC=-18142.92
PROMEDIO_GENERAL ~ AA1P + APE1P + ACD1P + AA2P + APE2P + ACD2P

                     Df Sum of Sq     RSS      AIC
<none>                             0.0000 -18142.9
+ ASISTENCIA_GENERAL  1    0.0000  0.0000 -18142.0
+ EDAD                1    0.0000  0.0000 -18141.1
- APE2P               1    9.4514  9.4514   -893.1
- AA2P                1   10.0127 10.0127   -877.5
- AA1P                1   11.2605 11.2605   -845.8
- APE1P               1   11.3741 11.3741   -843.1
- ACD1P               1   12.5933 12.5933   -815.6
- ACD2P               1   13.9688 13.9688   -787.6
```

Modelo ajustado

En el primer paso del ajuste, AIC tiene el menor valor por tanto se considera como mejor ajuste, eliminando las variables edad y asistencia general del modelo original, posteriormente se constata que las demás variables son necesarias para el modelo. Por tanto, el modelo final consta de seis variables regresoras (Ver Tabla 6).

lm(formula = PROMEDIO_GENERAL ~ AA1P + APE1P + ACD1P + AA2P + APE2P + ACD2P, data = papercor)

Tabla 5: Modelo de regresión lineal ajustado

(Intercept)	AA1P	APE1P	ACD1P	AA2P	APE2P	ACD2P
7.351e-15	1.667e-01	1.667e-01	1.667e-01	1.667e-01	1.667e-01	1.667e-01

Validación de supuestos del modelo ajustado

e=residuals(modelo_ajustado)

Se validan los supuestos del modelo ajustado en función del error residual.

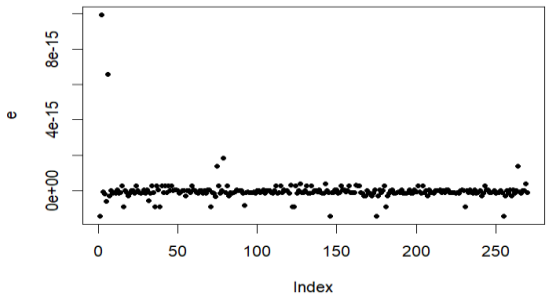
1. Independencia de errores

plot(e,pch=20): donde “e” guarda los datos de los errores residuales del modelo. El Gráfico 3 muestra, un tipo de patrón en los residuos para comprobar el supuesto se plantean las siguientes hipótesis:

H0: Los errores residuales son independientes

H1: Los errores no son independientes

Gráfico 3: Error residual del modelo



Para comprobar la hipótesis de independencia de errores aplicamos el test de Durbin-Watson. Los resultados iniciales no comprueban la H0 por lo que es necesario ajustar el modelo por segunda vez, insertando la variable EDAD (Ver Tabla 7).

modelo_ajustado=lm(formula = PROMEDIO_GENERAL ~ AA1P + APE1P + ACD1P + AA2P + APE2P + ACD2P, EDAD,data = papercor)

Tabla 6: Modelo ajustado_segundo

(Intercept)	AA1P	APE1P	ACD1P	AA2P	APE2P	ACD2P
-3.070e-14	1.667e-01	1.667e-01	1.667e-01	1.667e-01	1.667e-01	1.667e-01

En función del modelo ajustado se obtienen los siguientes valores para el test de Durbin-Watson

Tabla 7: Test de Durbin-Watson

DW = 2.222	p-value = 0.9652

El valor del p_valor es mayor que 0,05 por tanto, se comprueba la H0 que dice que los errores son independientes.

2. Normalidad

El siguiente supuesto a comprobar es la normalidad. La gráfica 4, muestra que los datos no siguen una distribución normal, los datos están sesgados al lado izquierdo de la gráfica.

Gráfico 4: Normalidad

Se utiliza entonces la prueba de Shapiro-Wilks para comprobar la normalidad

qqnorm(e, pch=20)

qqline(e)

https://doi.org/10.55204/pcc.v5i1.e96

Gráfico 5: Normalidad Q_Q Plot

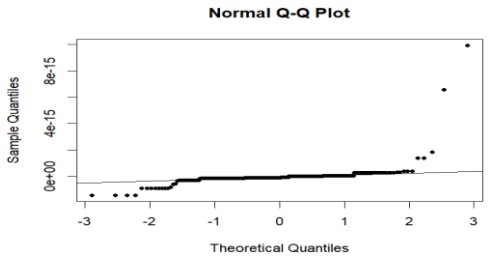
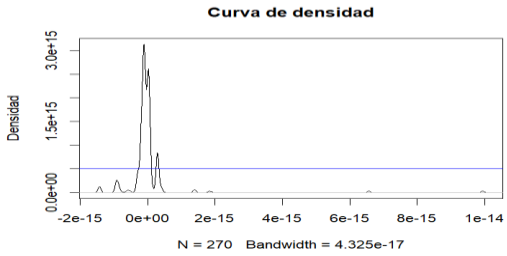


Gráfico 6: Curva de densidad de los errores residuales



El Gráfico 5, muestra que los datos tienen una linealidad creciente con ciertos datos atípicos, por tanto, podríamos afirmar que los cuartiles de la muestran son similares a los cuartiles teóricos. Por tanto, podría existir una posible distribución normal entre los errores. Por otra parte, el gráfico 6, indica la curva de densidad de los errores residuales.

Sin embargo, es necesario comprobar aplicando una prueba de normalidad para los errores basado en el test de Shapiro, planteamos la siguiente hipótesis:

H0: Los errores siguen una distribución de probabilidad normal

H1: Los errores no siguen una distribución de probabilidad normal

Tabla 8: Test de Shapiro:

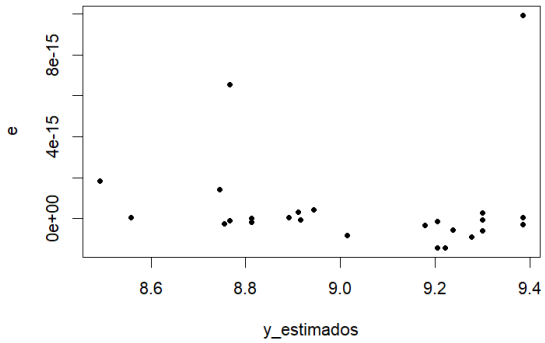
W = 0.30114	p-value < 2.2e-16

El p_valor del test de Shapiro, es menor a 0,05 por lo cual se acepta la H1, los datos no siguen una distribución de probabilidad normal.

3. Homocedasticidad

El gráfico 7, muestra en primera instancia que las variables no son constantes. Para comprobar usamos la Test (bptest).

prueba Breusch-Pagan Gráfico 7:Homocedasticidad



H0: Los errores tienen varianza constante

H1: Los errores no tienen varianza constante

bptest(modelo_ajustado)

Tabla 9:Prueba de Breusch-Pagan

BP = 4.0353	df = 6	p-value = 0.6719

El p_valor de la Prueba de Breusch-Pagan es mayor que 0,05 por tanto, se acepta la H0, los errores residuales si tienen varianza constante.

e. Evaluación del modelo

Para evaluar el modelo usamos la Prueba de Bondad de ajuste

Probamos la hipótesis generalx

H0: $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6$ *Todos los coeficientes de la variable regresora son* "0"

H1: $\beta_j = 0$. *Almenos* "1", *de los coeficientes de la variable regresora es distinto de* "0"

summary(modelo_ajustado)

```
Residuals:
       Min          1Q      Median          3Q         Max
-1.431e-15 -1.485e-16 -9.910e-17  4.880e-17  9.938e-15

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
(Intercept) -3.070e-14  4.055e-15 -7.572e+00 6.25e-13 ***
AA1P         1.667e-01  4.219e-16  3.950e+14  < 2e-16 ***
APE1P        1.667e-01  3.716e-16  4.485e+14  < 2e-16 ***
ACD1P        1.667e-01  5.588e-17  2.983e+15  < 2e-16 ***
AA2P         1.667e-01  2.109e-16  7.904e+14  < 2e-16 ***
APE2P        1.667e-01  1.450e-16  1.149e+15  < 2e-16 ***
ACD2P        1.667e-01  1.409e-16  1.183e+15  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.06e-16 on 263 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 4.748e+30 on 6 and 263 DF,  p-value: < 2.2e-16
```

Los resultados muestran que el p_valor es menor que 0,05; se acepta la H1 que menciona: *almenos* "1", *de los coeficientes de la variable regresora es distinto de* "0"

En la prueba de hipótesis individuales analizamos la columna de los Pr (>| t |). Si los valores son mayores, que 0,05; no son significativos, caso contrario son significativos para el modelo. Para el caso de estudio todas las variables son menores a 0,05 por consiguiente son significativas para el modelo.

Finalmente, el R2 (Adjusted R-squared) es igual a 1, por lo tanto, se puede afirmar que el modelo de regresión lineal ajustado, explica en un 100% de la variabilidad de toda la base de datos.

https://doi.org/10.55204/pcc.v5i1.e96

3. RESULTADOS Y DISCUCIÓN

Se realizó, previamente, un análisis descriptivo con respecto a las variables cuantitativas, reflejados en la Tabla 2

Tabla 10:Estadística Descriptiva

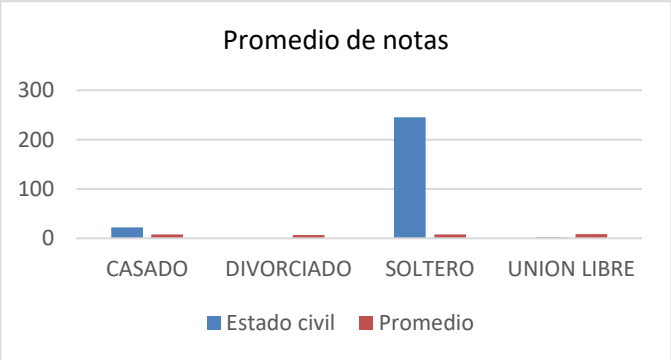
	PROMEDIO_GENERAL	EDAD	AA1P	APE1P	ACD1P	AA2P	APE2P	ACD2P	ASISTENCIA_GENERAL
Media	8.0	21.2	8.7	8.9	6.7	8.3	8.3	7.3	93.1
Mediana	8.4	20.4	9.0	9.5	6.9	8.8	8.9	8.0	100.0
Desviación estándar	1.5	3.3	1.7	1.6	1.9	1.8	1.9	2.2	15.4
Max	9.8	40.7	10.0	10.0	10.0	10.0	10.0	10.0	100.0
Min	0.0	17.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0

En general, los estudiantes tienen un promedio general de 8.0 en la asignatura de estadística, la media de las edades es de 21 años, el promedio del aprendizaje autónomo en el primero y segundo parcial es de 8.7 y 8.3, respectivamente, en tanto que en el aprendizaje practico experimental es de 9.5 y 8.9 en el primero y segundo parcial, para el aprendizaje en contacto con el docente el promedio es de 6.9 y 8.0.

Del gráfico 8 se deduce que los promedios de las personas que tiene estado civil “Unión libre”, es el más bajo en comparación con aquellos que son solteros, casados o divorciados.

Gráfico 8: Notas promedio por estado civil

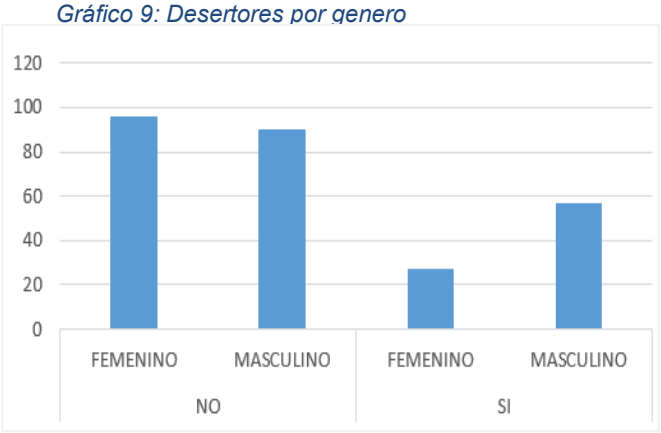
	Estado civil	Promedio
Casado	22	8,0
Divorciado	1	6,3
Soltero	245	8,0
Unión libre	2	8,7
Total	270	



Fuente: Elaboración propia en Excel

Gráfico 9: Desertores por genero

Género - Desertor	Valor
NO	186
Femenino	96
Masculino	90
SI	84
Femenino	27
Masculino	57
Total	270



Fuente: Elaboración propia en Excel

https://doi.org/10.55204/pcc.v5i1.e96

El gráfico 9, representa los estudiantes que han desertado según su género, de ellos 57 son de género masculino y 27 femenino.

4. CONCLUSIONES

Para ejecutar la técnica multivariante de regresión lineal múltiple, necesario verificar los supuestos, este procedimiento permite afirmar la robustez del análisis estadístico. Las técnicas multivariadas son útiles para analizar el rendimiento académico de los alumnos.

Puesto que el rendimiento académico está directamente ligado con la deserción estudiantil, los resultados muestran, que las estudiantes de género femenino tienen menor tendencia a la deserción que los hombres, por tanto, la deserción en el género femenino es menor que en el masculino.

La mayor cantidad de outsiders se encontró en la variable APE1P, misma que tiene el promedio más bajo en relación a las demás variables predictoras. Existe una elevada dispersión en la variable edad, y aprendizaje en contacto con el docente del segundo parcial, en relación al valor medio.

Las variables predictores que inciden directamente en el rendimiento académico de la asignatura de estadística son AA1P, APE1P, ACD1P, AA2P, APE2P, ACD2P, EDAD.

Se comprobaron los supuestos de independencia de errores aplicando la prueba de Durbin-Watson. A pesar que los resultados iniciales no comprueban la H0, fue necesario ajustar el modelo, insertando la variable EDAD, para la comprobación con un p-value igual a 0.9652 mayor superior a 0,05.

Acerca de la normalidad se comprueba que los residuos no siguen una distribución normal, los datos están sesgados al lado izquierdo de la gráfica. Los datos tienen una linealidad creciente con ciertos datos atípicos, por tanto, podríamos afirmar que los cuartiles de la muestran son similares a los cuartiles teóricos.

Se comprueba la homocedasticidad con la prueba de Breusch-Pagan. El p_valor de la prueba es mayor que 0,05 por tanto, los errores residuales si tienen varianza constante con relación al modelo ajustado.

Se evalúa el modelo usando la prueba de bondad de ajuste en donde se determina que almenos "1", de los coeficientes de la variable regresoras es distinto de "0"

El modelo creado, generó una función de regresión lineal múltiple que predice el rendimiento académico en función del promedio general de un estudiante. Se realiza una breve descripción de modelo obtenido, se verifica el cumplimiento de supuestos y los resultados obtenidos posterior a la aplicación del modelo. Se determinan los supuestos, con el propósito de conocer si la solución encontrada a través de la aplicación de la regresión lineal múltiple en la generación del modelo predictivo es estable.

AGRADECIMIENTOS

A la Universidad Politécnica Estatal del Carchi.

FINANCIACIÓN

Los autores no recibieron financiación para el desarrollo de la presente investigación.

CONFLICTO DE INTERESES

Los Autores declaran que no existe conflicto de intereses

CONTRIBUCIÓN DE AUTORÍA

Participar activamente en:	*Cuji B.*	*Gavilánes W..*
Conceptualización	X	
Análisis formal	X	
Adquisición de fondos	X	
Investigación		X
Metodología	X	
Administración del proyecto	X	X
Recursos		X
Redacción –borrador original	X	X
Redacción –revisión y edición		
La discusión de los resultados	X	
Revisión y aprobación de la versión final del trabajo.	X	X

REFERENCIAS

Amín, A., & Amezcua, L. (2021). Construcción de un modelo predictivo para determinar el rendimiento académico de los estudiantes del colegio de estudios científicos y tecnológicos del estado de Michoacán. *Ciencia Latina Revista Científica Multidisciplinar*, *5*(5), 7709–7749. https://doi.org/10.37811/cl_rcm.v5i5.872

Cabrera, F., Verdugo, M. E., Cabrera, H., Escudero, M., & Franco, M. (n.d.). *Rendimiento académico universitario, según el modelo de bachillerato por especialidades y el Bachillerato General Unificado del Ecuador, Estudio de caso de la Universidad de Cuenc 2012-2018*.

Campo, N. M. S. del, & Matamoros, L. Z. (2020). Técnicas estadísticas multivariadas para el estudio de la causalidad en Medicina. *Revista de Ciencias Médicas de Pinar Del Río*, *24*. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1561-31942020000200287

Castelazo, I. Y. L., Mercado, A. G., Juárez, N. Y. L., Pérez, N. G. G., & Juárez, M. I. S. (2022). Modelo Predictivo del Rendimiento Académico de Estudiantes de Ingeniería Química en el Área de Matemáticas. *Journal of Engineering Research*, *2*(1), 2–10. https://doi.org/10.22533/at.ed.317212210018

Clifford, B., & Taylor, R. (2014). Bioestadística. In PEARSON (Ed.), *Bioestadística* (Primera Ed). https://doi.org/10.22201/fesz.9786070261015p.2014

Collazos, A. C., Quintero Medina, M. V., & Trujillo Caicedo, K. N. (2021). Determinantes del Rendimiento Academico de la prueba Saber 11 Durante el periodo 2014-2019 en Colombia. *Panorama*, *15*(29), 103–126. https://doi.org/10.15765/pnrm.v15i29.1723

Dorta Guerra, R., Marrero, I., Abdul-Jalbar, B., Trujillo González, R., & Torres Darias, N. (2019). Un modelo predictivo del rendimiento académico a partir de las calificaciones de Bachillerato y PAU. *De Los Procesos de Cambio Al Cambio Con Sentido*, 119–136. https://doi.org/10.25145/b.innovaull.2019.009

Fernández, Y. O. (2021). Variables académicas que influyen en el rendimiento académico de los estudiantes universitarios. *Investigación Educativa*, *15*, 165–179. http://revistasinvestigacion.unmsm.edu.pe/index.php/educa/article/viewFile/6473/5692

Granados, R. M. (2016). *Modelos de regresión lineal múltiple*.

https://www.ugr.es/~montero/matematicas/regresion_lineal.pdf

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (2019). Análisis multivariante. In *Prentice Hall* (Quinta Edi, Vol. 53, Issue 9). https://doi.org/10.1017/CBO9781107415324.004

Herreras, E. (2019). Estudio Predictivo del Rendimiento Matemático en PISA 2012: Enfoque de Aprendizaje Frente a la Atribución del Fracaso. *Revista Iberoamericana de Diagnóstico y Evaluación – e Avaliação Psicológica*, *52*(3), 156–171. https://doi.org/10.21865/ridep52.3.12

Lamas, H. A. (2014). Sobre el rendimiento escolar School. *Propósitos y Representaciones*, *3*(1), 141–160. https://doi.org/10.4135/9781483328416.n9

Naal, J. E., Pacheco, A. del Á., & López, C. D. J. (2022). Relacion entre el rendimiento académico y los resultados del rendimiento académico de alumnos de gastronomía. *Ciencia Latina Revista Científica Multidisciplinar*, *6*(4), 83–98. https://doi.org/10.37811/cl_rcm.v6i4.2517

Ortega, E., Solís, J. F., & Martínez, D. (2022). Deserción de los alumnos de Ingeniería en Computación: Un Modelo Predictivo. *Journal of Engineering Research*, *2*(5), 2–8. https://doi.org/10.22533/at.ed.317252208046

Rico Páez, A., Gaytán Ramírez, N. D., & Sánchez Guzmán, D. (2019). Construcción e implementación de un modelo para predecir el rendimiento académico de estudiantes universitarios mediante el algoritmo Naïve Bayes. *Diálogos Sobre Educación*, *19*, 1–18. https://doi.org/10.32870/dse.v0i19.509

Tortolero Portugal, R., Figueroa González, E. G., & Villareal Solís, F. M. (2020). Modelo de Regresión Lineal Múltiple de la Gestión del Conocimiento, con la Cultura Organizacional, el Liderazgo y las Tecnologías de la Información y la Comunicación, en Trabajadores de una Empresa de la Cd. de Durango, Durango, México. *Hitos de Ciencias Económico Administrativas*, *26*(76), 266–284. https://doi.org/10.19136/hitos.a26n76.4089

Vázquez, S. R., Enrique, J., & Vázquez, R. (2022). Factores psicológicos predominantes y rendimiento académico de alumnos de estadística, universidad nacional de Piura, 2019. *Revista Cuántica*, *1*, 14–25. https://doi.org/10.56747/rcq.v1i1.21